Insper

Seleção de modelos, regularização e regressão logística

Aula 3

Magno Severino
PADS - Modelos Preditivos
08/04/2021

Objetivos de aprendizagem

Ao final dessa aula você deverá ser capaz de

- compreender e aplicar técnicas de seleção de variável;
- relacionar técnicas de regularização com o trade-off viés-variância;
- ajustar, definir hiperparâmetros e aplicar modelos com técnicas de regularização;
- comparar modelos de regressão linear e regularizados.
- Conceituar a regressão logística;
- Avaliar perfomance de um modelo de classificação.

Seleção de modelos

- **Seleção de subconjuntos:** considera um subconjunto das p preditoras.
- **Regularização:** ajusta-se um modelo com as p preditoras e os coeficientes estimados são encolhidos em direção a zero. Essa abordagem reduz a variância.
- **Redução de dimensão:** considera a utilização de uma combinação das p preditoras numa dimensão M tal que $M < p$.

Critérios

- $\mathrm{C}_p = \frac{1}{n}(\mathrm{RSS} + 2p\hat{\sigma}^2)$,
- Akaike Information Criteria: $\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2p\hat{\sigma}^2\right)$,
- Bayesian Information Criteria: $\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)p\hat{\sigma}^2\right)$,

em que p é o número de preditoras utilizadas no modelo e $\hat{\sigma}^2$ é uma estimativa da virância do erro ϵ baseado em

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Best subset selection

- Seja $\mathcal{M}_0$ o modelo nulo (aquele que prevê apenas pela média de Y).
- Para $k = 1, \ldots, p$,
    - ajuste todos os $\binom{p}{k} = \frac{p!}{(p-k)!k!}$ modelos com k preditoras;
    - selecione o melhor entre todos os $\binom{p}{k}$ modelos ajustados e denote-o por $\mathcal{M}_k$.
    - o melhor modelo pode ser definido de acordo com RSS ou R^2.
- Selecione o melhor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.
- **Problema?**

Stepwise

Para contornar o problema do número de modelos do método *best subset selection*, as abordagens *stepwise* exploram um espaço restrito de modelos.

Número de variáveis	Best subset	Stepwise
2	4	4
4	16	11
8	256	37
16	65536	137
32	4294967296	529

Forward stepwise selection

- Seja $\mathcal{M}_0$ o modelo nulo.
- Para $k = 0, \ldots, p-1$
    - considere todos os $p-k$ modelos que aumentam as preditoras no modelo $\mathcal{M}_k$ em uma preditora.
    - escolha o melhor modelo entre os $p-k$ modelos e denote por $\mathcal{M}_{k+1}$.
    - novamente, o melhor modelo pode ser definido como a menor RSS ou maior R^2.
- Selecione o melhor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.

Esse método pode ser aplicado para os cenários de alta dimensão $(n < p)$. No entanto, para esses casos, é possível construir os modelos $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}$. Pois o método dos mínimos quadrados não possui solução única para os casos em que $p \geq n$.

Backward stepwise selection

- Seja $\mathcal{M}_0$ o modelo nulo.
- Para $k = p, p-1, \ldots, 1$
  - considere todos os $p - k$ modelos que contenham todas as preditoras no modelo $\mathcal{M}_k$ menos uma, para um total de $k - 1$ preditoras.
  - escolha o melhor modelo entre os $p - k$ modelos e denote por $\mathcal{M}_{k-1}$.
  - novamente, o melhor modelo pode ser definido como a menor RSS ou maior R^2.
- Selecione o melhor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ utilizando validação cruzada para o erro de previsão, C_p, AIC, BIC ou R^2 ajustado.

Dados Credit [1]

- **ID**: id
- **Income**: renda (em $10,000)
- **Limit**: limite de crédito
- **Rating**: rating de crédito
- **Cards**: número de cartões de crédito
- **Age**: idade em anos
- **Education**: anos de escolaridade
- **Gender**: Male / Female
- **Student**: Yes / No
- **Married**: Yes / No
- **Ethnicity** : African American / Asian / Caucasian
- **Balance**: saldo médio do cartão de crédito em $

[1] Fonte: livro An Introduction to Statistical Learning with Applications in R.

Dados Credit

```r
library(ISLR)
data(Credit)
```

ID	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity
118	91.362	9113	626	1	47	17	Male	No	Yes	Asian
352	61.62	5140	374	1	71	9	Male	No	Yes	Caucasian
187	36.472	3806	309	2	52	13	Male	No	No	African American
93	30.733	2832	249	4	51	13	Male	No	No	Caucasian
337	32.856	5884	438	4	68	13	Male	No	No	Caucasian

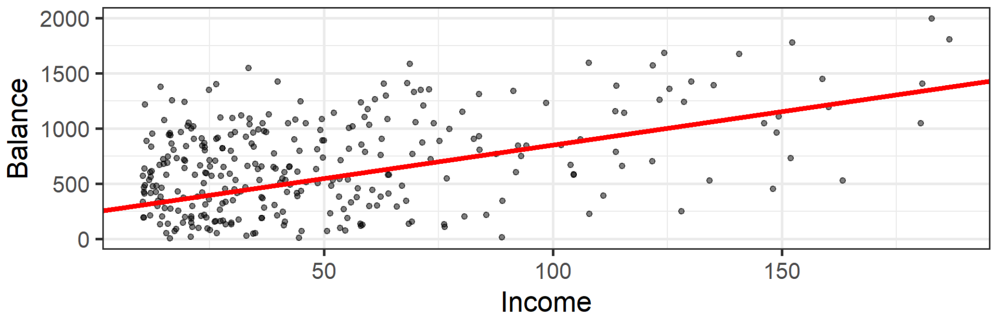
Como representar a variável *Student*?

- Valores que **Student** pode assumir: Yes / No.
- Faz sentido escrever o modelo abaixo?

$$Balance = \beta_0 + \beta_1 Income + \beta_2 Student + \epsilon.$$

- Alternativa:

$$\mathbb{I}(Student) = \begin{cases} 1, & \text{se } Student = Yes \\ 0, & \text{caso contrário.} \end{cases}$$



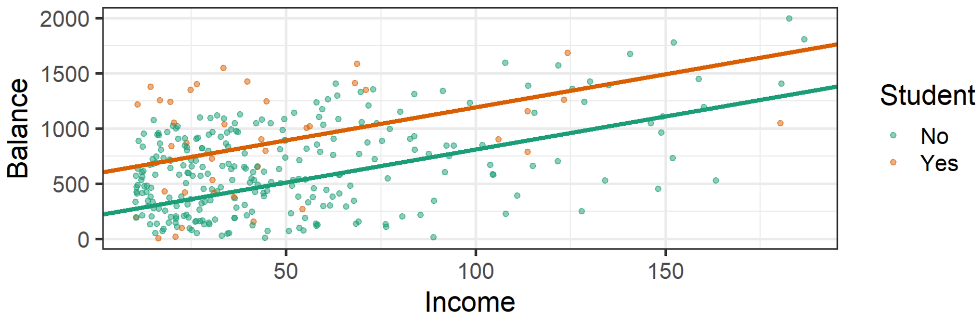
Como representar a variável *Student*?

- Valores que **Student** pode assumir: Yes / No.
- Faz sentido escrever o modelo abaixo?

$$Balance = \beta_0 + \beta_1 Income + \beta_2 Student + \epsilon.$$

- Alternativa:

$$\mathbb{I}(Student) = \begin{cases} 1, & \text{se } Student = Yes \\ 0, & \text{caso contrário.} \end{cases}$$



Prática R

- Faça uma análise exploratória dos dados (*exploratory data analysis - EDA*).
- Quais variáveis você acredita que mais se relacionam com `Balance`?

Dados Credit

```
fit <- lm(Balance ~ ., data = Credit[,-1])

summary(fit)$coefficients
```

```
##                       Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)        -479.2078706 35.77393717 -13.3954468  6.730600e-34
## Income               -7.8031018  0.23423191 -33.3135727 7.372312e-116
## Limit                 0.1909067  0.03277862   5.8241238  1.205974e-08
## Rating                1.1365265  0.49089445   2.3152157  2.112213e-02
## Cards                17.7244836  4.34103295   4.0830106  5.401200e-05
## Age                  -0.6139088  0.29398941  -2.0882005  3.743127e-02
## Education            -1.0988553  1.59795129  -0.6876651  4.920746e-01
## GenderFemale        -10.6532477  9.91399990  -1.0745660  2.832368e-01
## StudentYes          425.7473595 16.72258016  25.4594300  8.854521e-85
## MarriedYes           -8.5339006 10.36287466  -0.8235071  4.107256e-01
## EthnicityAsian       16.8041792 14.11906302   1.1901767  2.347047e-01
## EthnicityCaucasian   10.1070252 12.20992331   0.8277714  4.083088e-01
```

Observação

Note que, ao utilizar o *best subset*, é possível que aconteça a situação seguinte.

# de variáveis	Best subset	Forward stepwise
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	cards, income, student, limit	rating, income, student, limit

Os três primeiros modelos em cada coluna são idênticos, já o quarto é diferente.

No método *forward stepwise*. uma variável que aparece no primeiro moedelo fará parte de todos os modelos até o último passo (modelo final).

Stepwise

Forward

```
library(MASS)
fit <- lm(Balance ~ 1, data = Credit[,-1])
stepAIC(fit, direction = "forward",
  scope = list(lower = ~ 1,
               upper = ~ Income + Limit + Rating + Cards + Age +
                 Education + Gender + Student + Married +
                 Ethnicity))
```

Backward

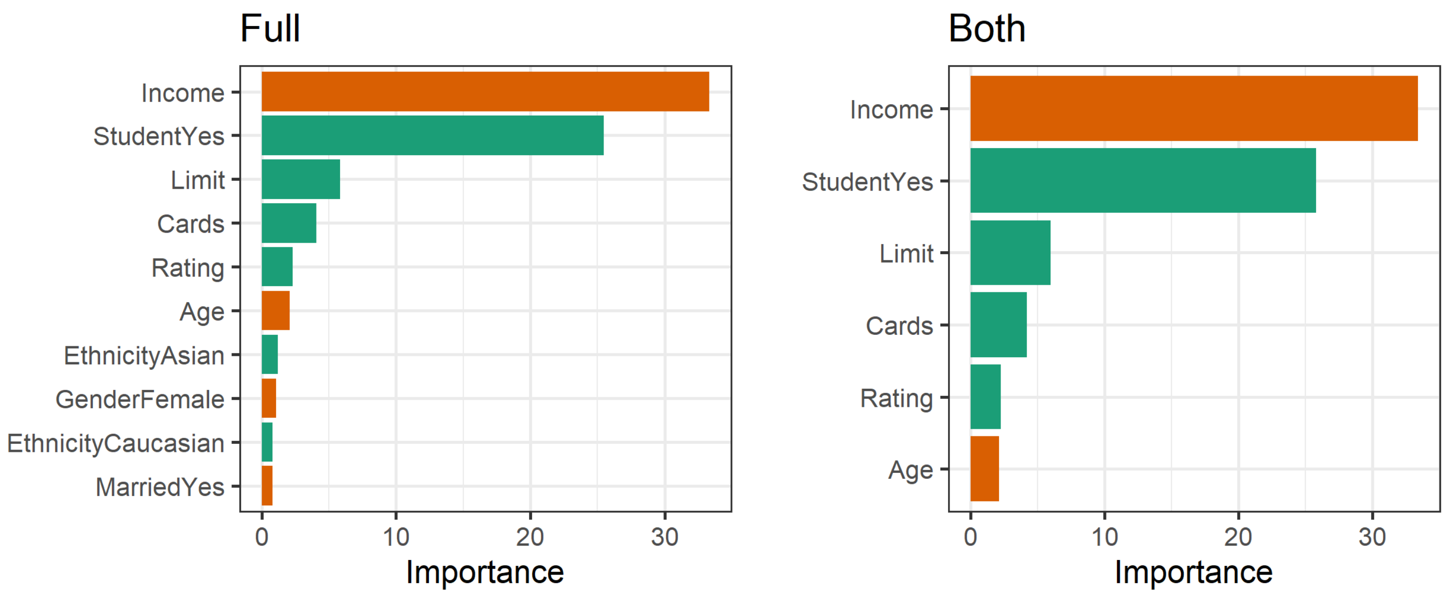
```
fit <- lm(Balance ~ ., data = Credit[,-1])
stepAIC(fit, direction = "backward")
```

Both

```
stepAIC(fit, direction = "both")
```

Importância de variáveis

```r
library(patchwork)
library(vip)

fit1 <- lm(Balance ~ ., data = Credit[, -1])
fit2 <- lm(Balance ~ Income + Limit + Rating + Cards + Age +
                     Student, data = Credit[, -1])

g1 <- vip(fit1, mapping = aes(fill = Sign)) + labs(title = "Full")
g2 <- vip(fit2, mapping = aes(fill = Sign)) + labs(title = "Both")

g1 + g2
```



Métodos de encolhimento

Regressão Ridge

No modelo de regressão linear, o objetivo é encontrar $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ que minimizam

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Agora, vamos considerar um termo de penalização para a expressão acima. Assim,

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 .$$

Qual o impacto que este termo de penalização causa no RSS? O que acontece se $\lambda = 0$? E se $\lambda \to \infty$?

Minimizar a quantidade acima é equivalente à resolver o seguinte problema de otimização

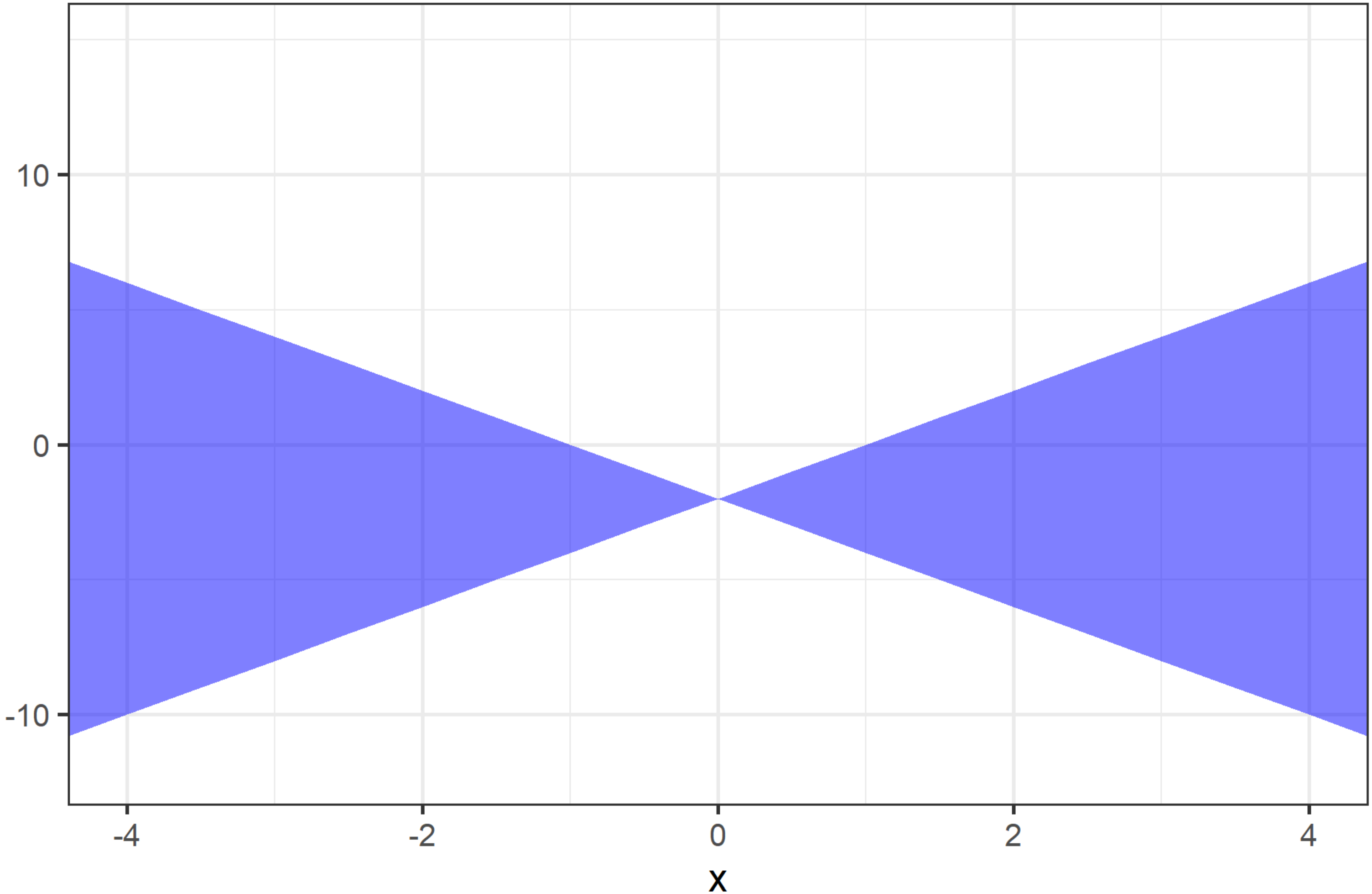
$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{sujeito a} \quad \sum_{j=1}^{p} \beta_j^2 \leq s .$$

Existe uma relação entre λ e s.

Note que β_0 **não** é regularizado.

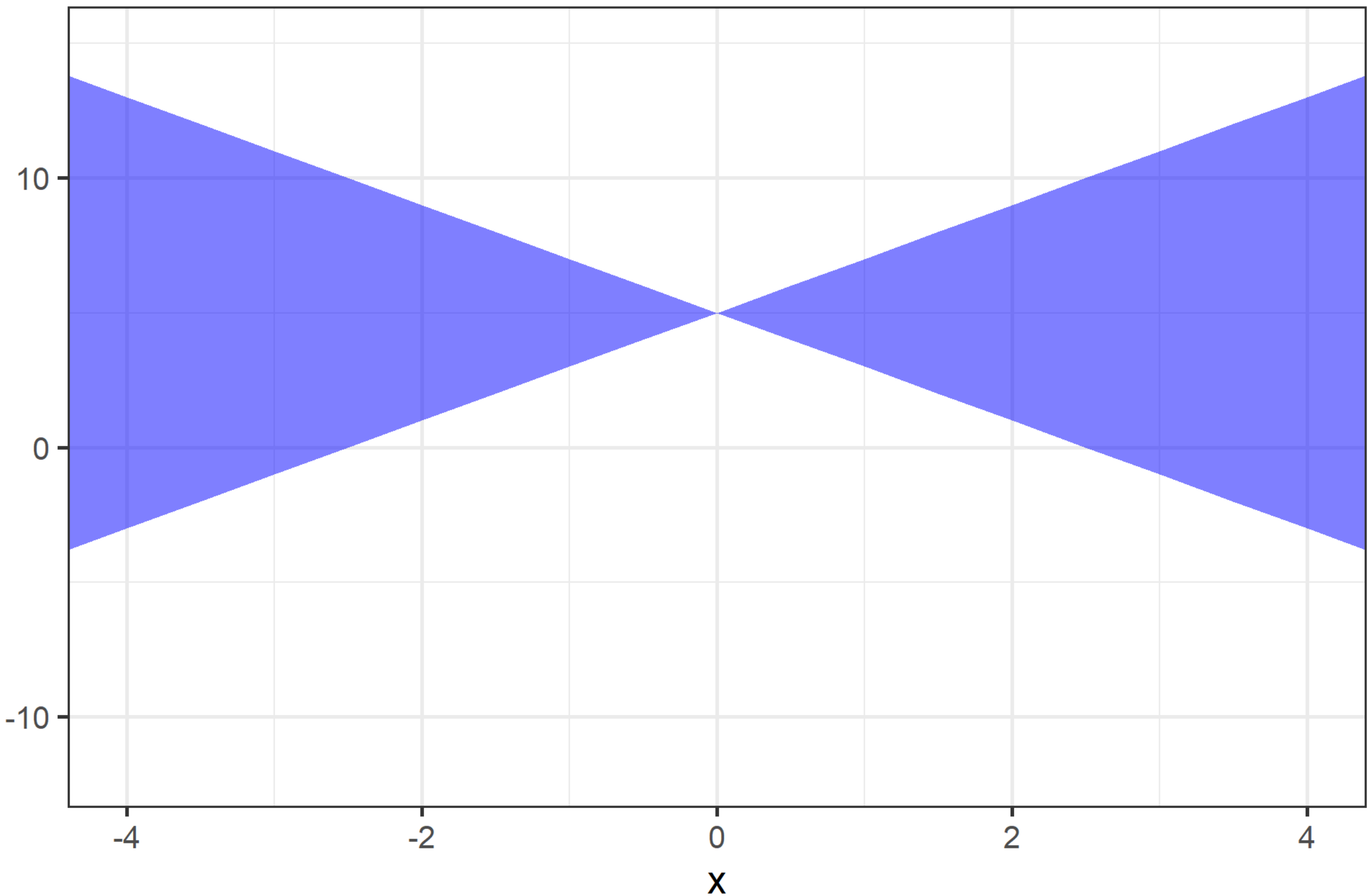
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, i.e. $-2 \leq \beta_1 \leq 2$.



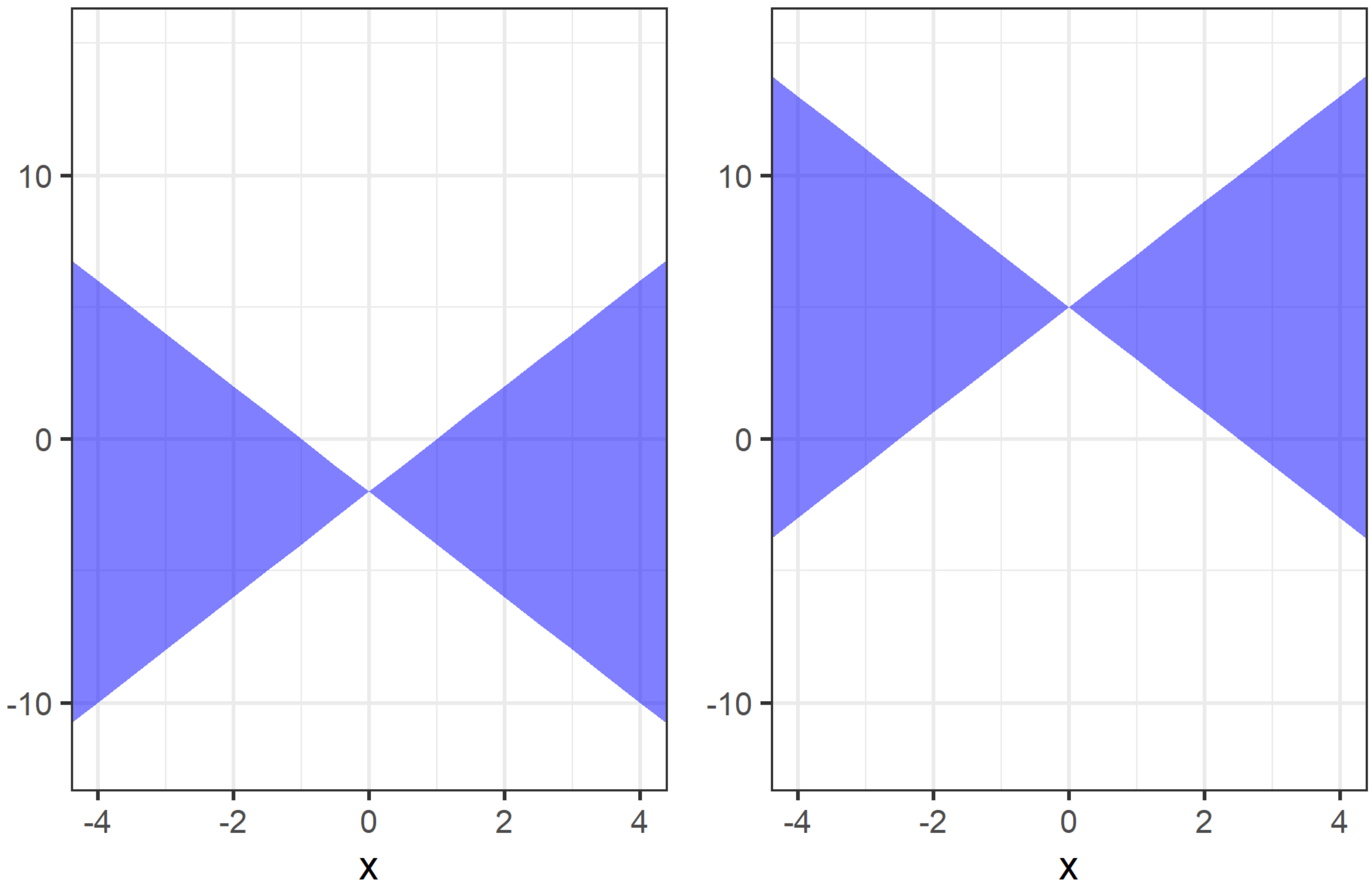
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



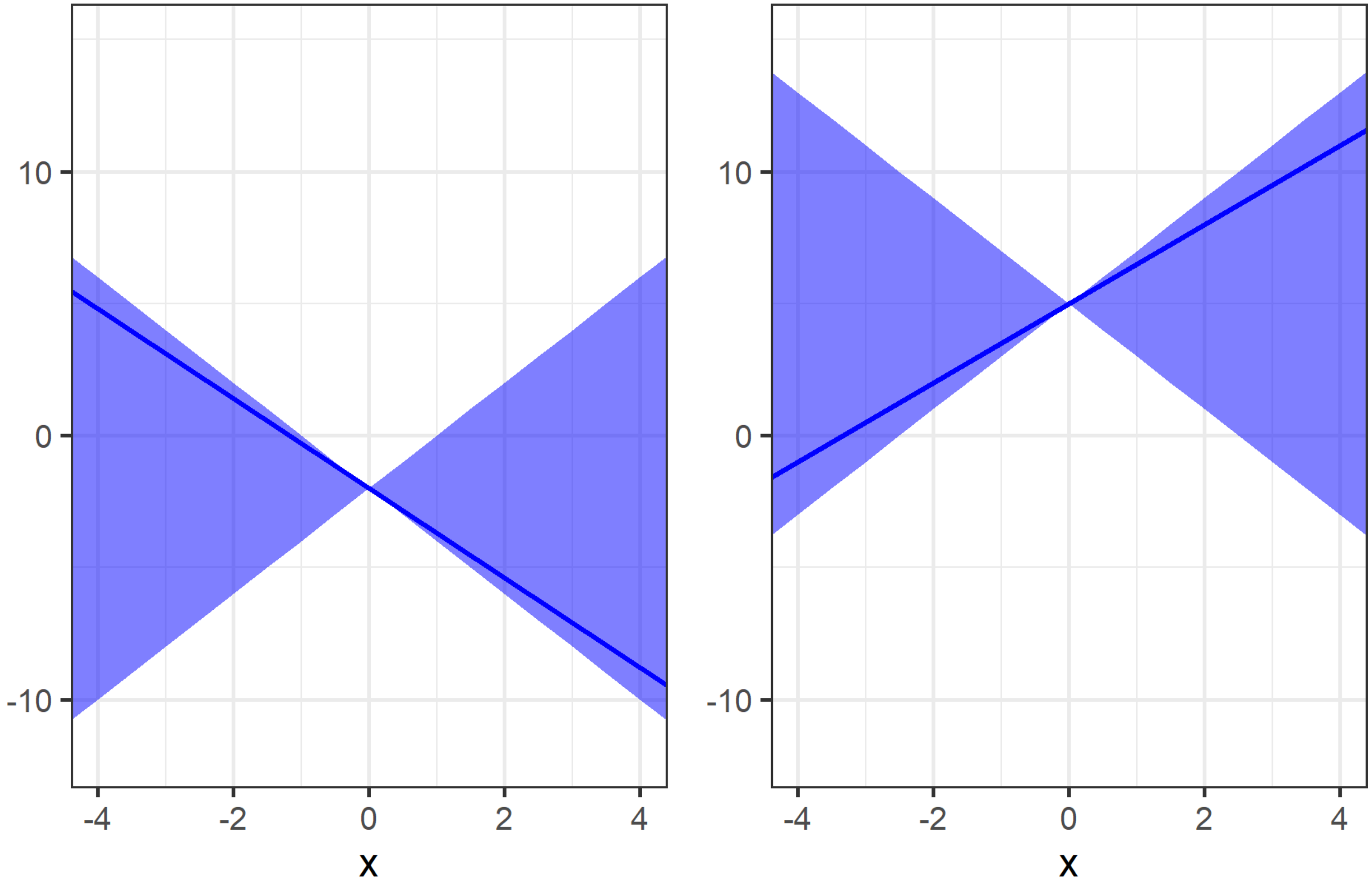
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



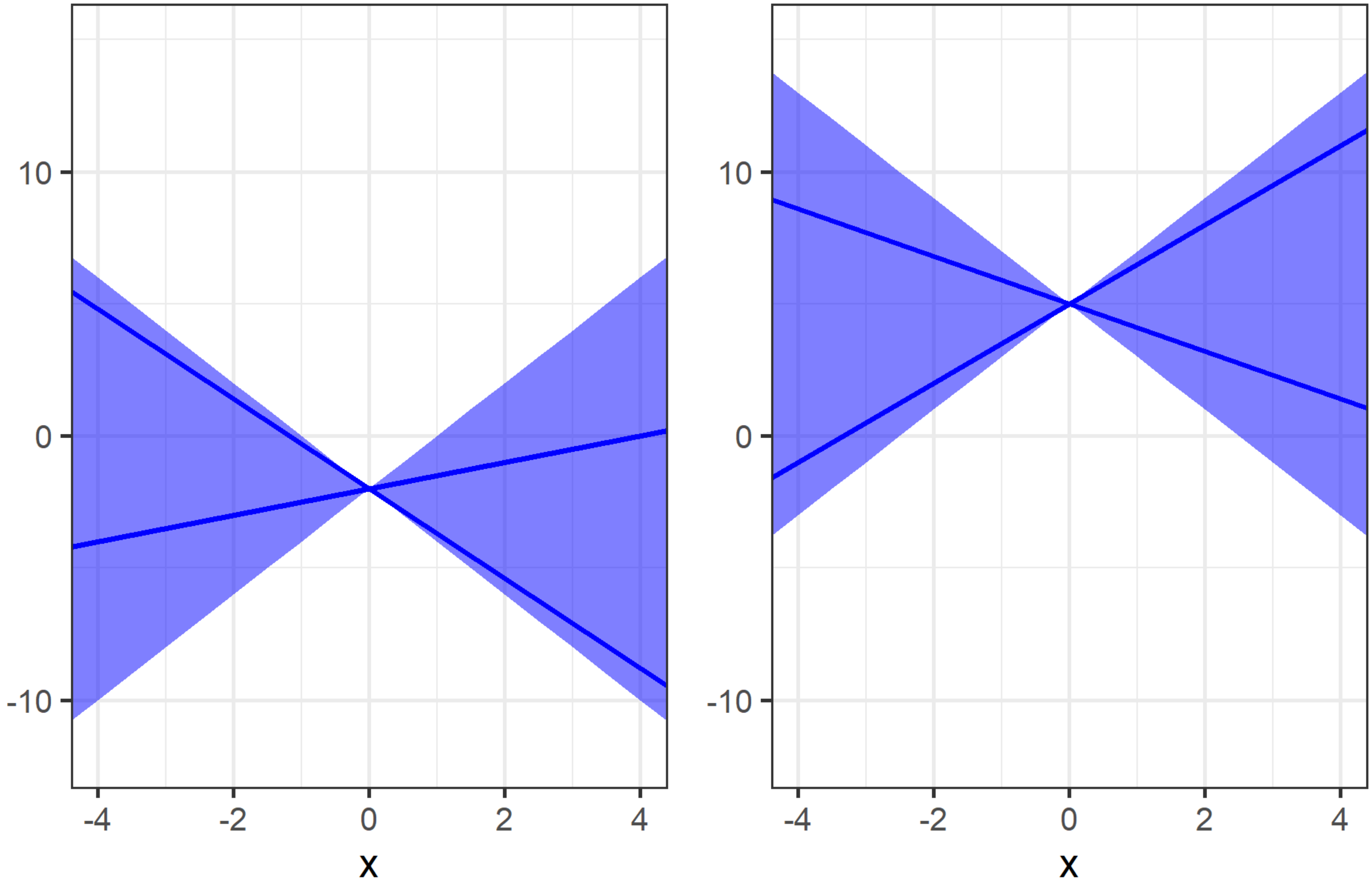
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



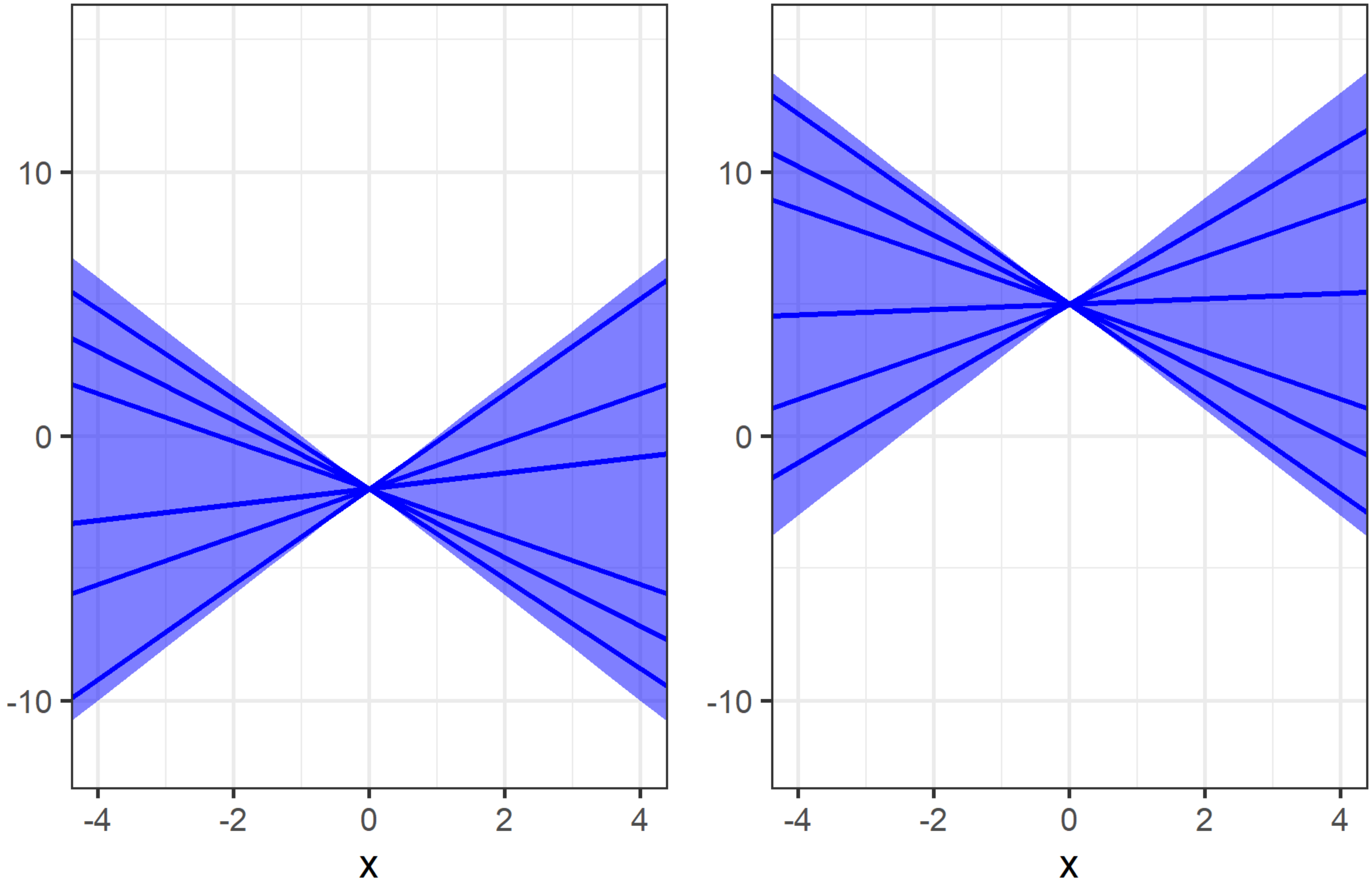
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



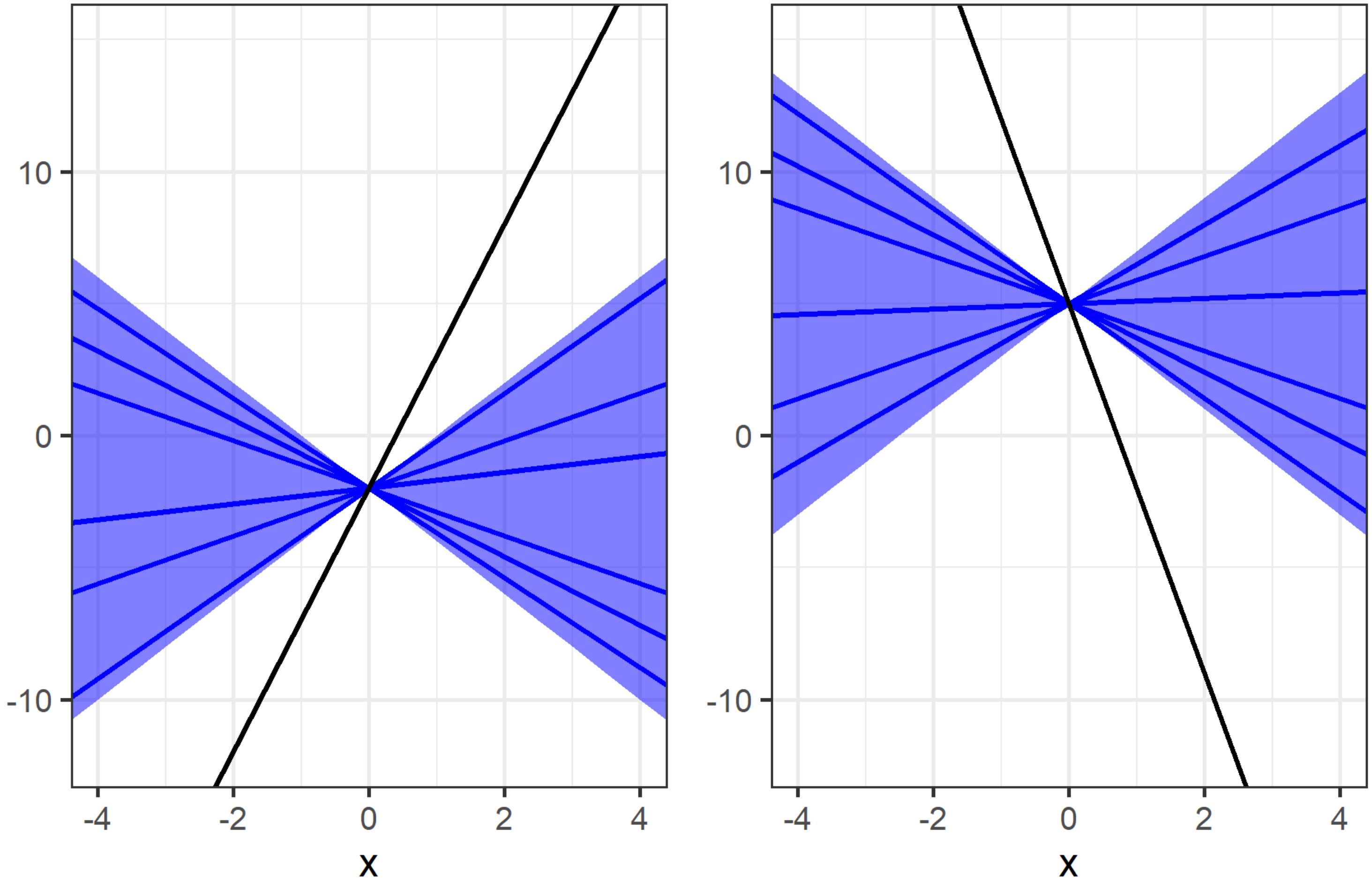
Regressão Ridge

Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



Regressão Ridge

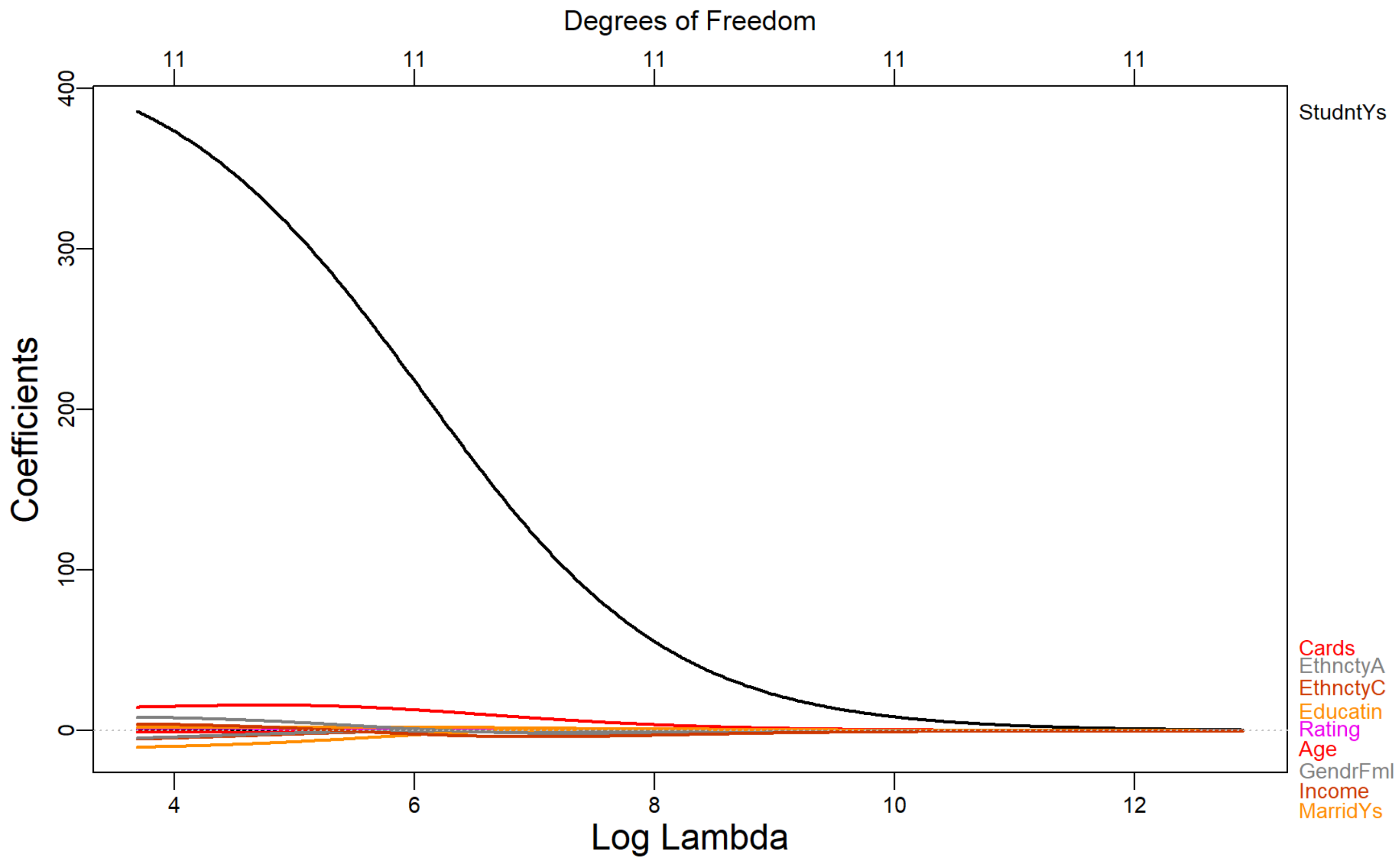
Considere o caso em que $Y = \beta_0 + \beta_1 X_1$. Se $s = 2$ então $|\beta_1| \leq 2$, então $-2 \leq \beta_1 \leq 2$.



Regressão Ridge

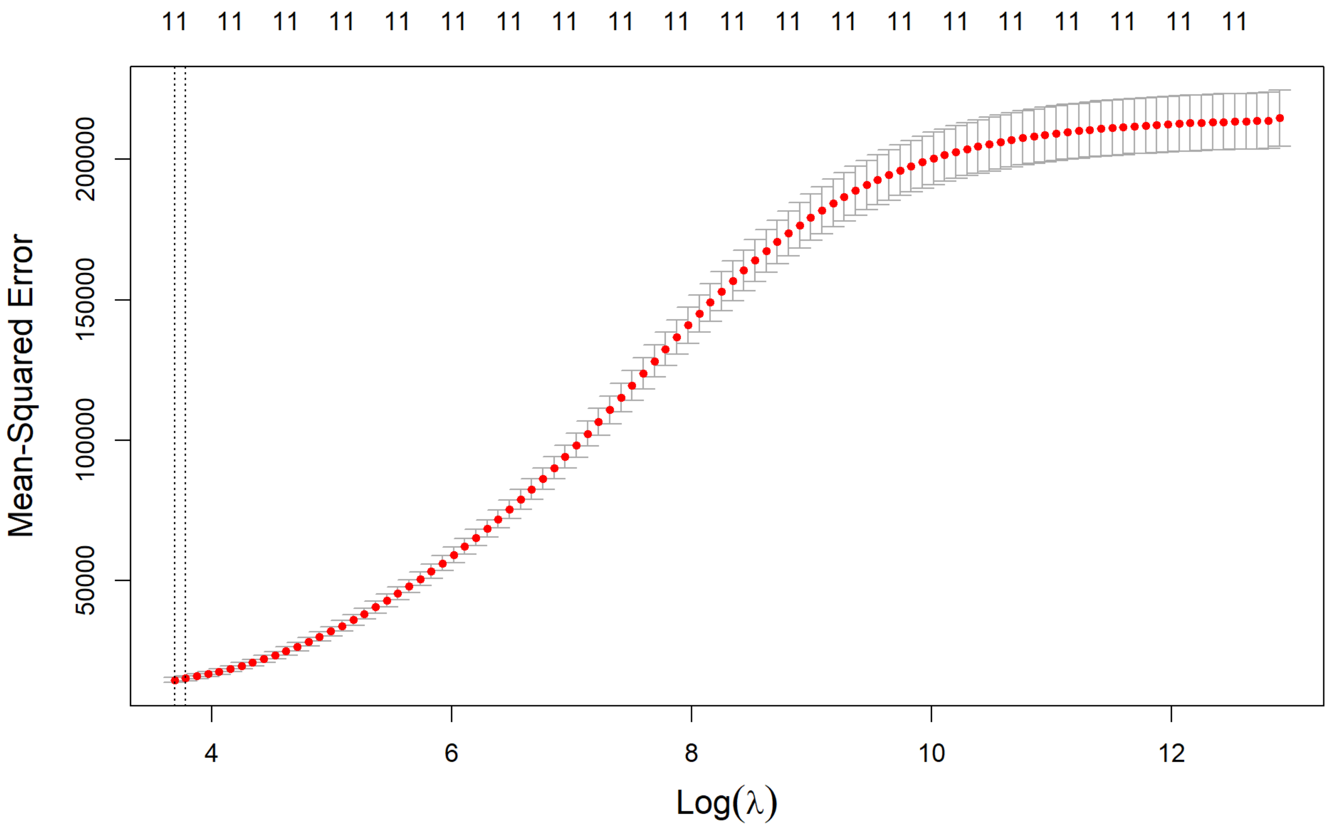
```r
library(ISLR)    # base de dados
library(glmnet)  # LASSO, ridge e elasticnet
library(plotmo)  # gráficos

X <- model.matrix(Balance ~ ., data = Credit[,-1])[,-1]
y <- Credit$Balance

set.seed(12)
idx <- sample(nrow(Credit), size = .75*nrow(Credit),
              replace = FALSE) # indice treinamento

ridge <- glmnet(X[idx,], y[idx], alpha = 0, nlambda = 500)

plot_glmnet(ridge, lwd = 2, cex.lab = 1.3)

ridge$a0
ridge$beta
ridge$lambda
```

Regressão Ridge



Regressão Ridge - escolha de λ

Vamos determinar λ através da validação cruzada.



Esse gráfico apresenta a estimativa do erro e o desvio-padrão. Essa função utiliza 10-folds como padrão. Os números na parte superior indicam quantos coeficientes são diferentes de zero.

Regressão Ridge - resultados

```r
y_ridge <- predict(ridge, newx = X[-idx,],
                   s = cv_ridge$lambda.1se)

tab <- tibble(metodo = c("lm", "ridge", "lasso", "elastic"),
              mse = NA)

tab$mse[tab$metodo == "ridge"] <- mean((y[-idx] - y_ridge)^2)

fit_lm <- lm(Balance ~ ., Credit[idx, -1])
y_lm <- predict(fit_lm, Credit[-idx,])
tab$mse[tab$metodo == "lm"] <- mean((y[-idx] - y_lm)^2)

tab
```

```
## # A tibble: 4 x 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lm      11205.
## 2 ridge   14606.
## 3 lasso      NA
## 4 elastic    NA
```

Regressão LASSO

A regressão LASSO (Least Absolute Shrinkage and Selection Operator), consdiera a seguinte penalização

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

O que acontece se $\lambda = 0$? E se $\lambda \to \infty$?

Minimizar a quantidade acima é equivalente à resolver o seguinte problema de otimização

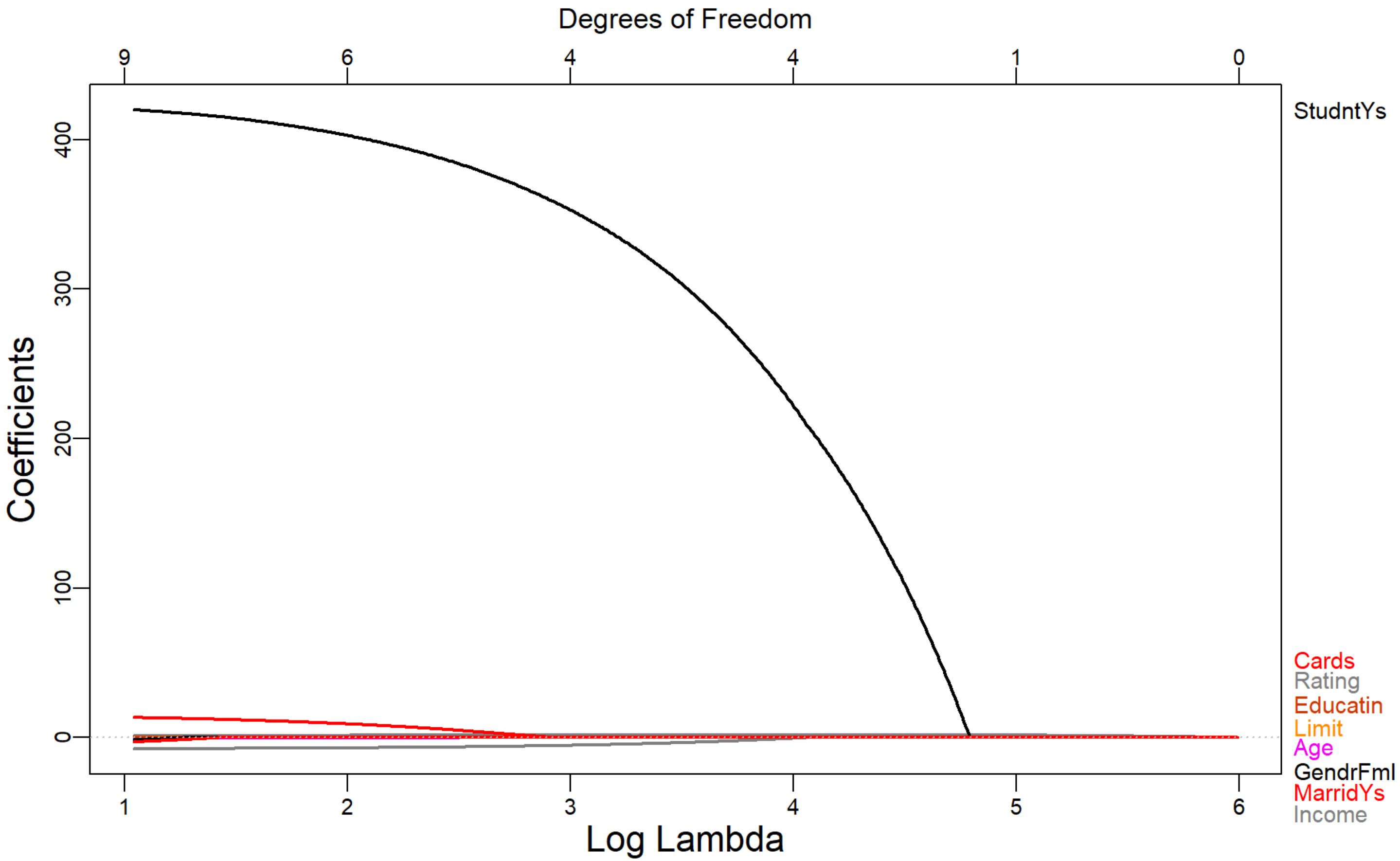
$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{sujeito a} \quad \sum_{j=1}^{p} |\beta_j| \leq s.$$

Essa penalização é comumente denotada como ℓ_1, pois a norma ℓ_1 de um vetor $\boldsymbol{\beta}$ é dada por $\ell_1 = \sum_j |\beta_j|$.

Existe uma relação entre λ e s. Note que β_0 não é regularizado.

Regressão LASSO

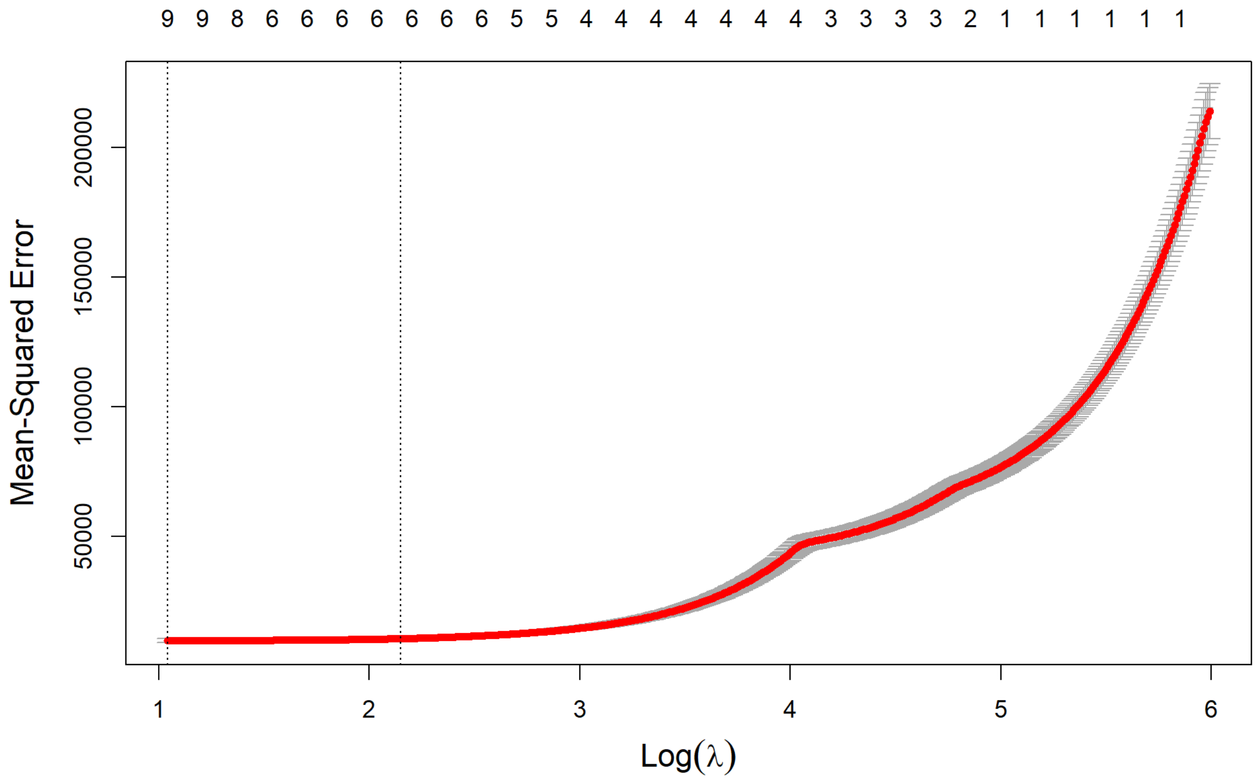
```r
lasso <- glmnet(X[idx,], y[idx], alpha = 1, nlambda = 1000)
plot_glmnet(lasso, lwd = 2, cex.lab = 1.3, xvar = "lambda")
```



Regressão LASSO - escolha de λ

Vamos determinar λ através da validação cruzada.

```
cv_lasso <- cv.glmnet(X[idx,], y[idx], alpha = 1,
                      lambda = lasso$lambda)
plot(cv_lasso, cex.lab = 1.3)
```

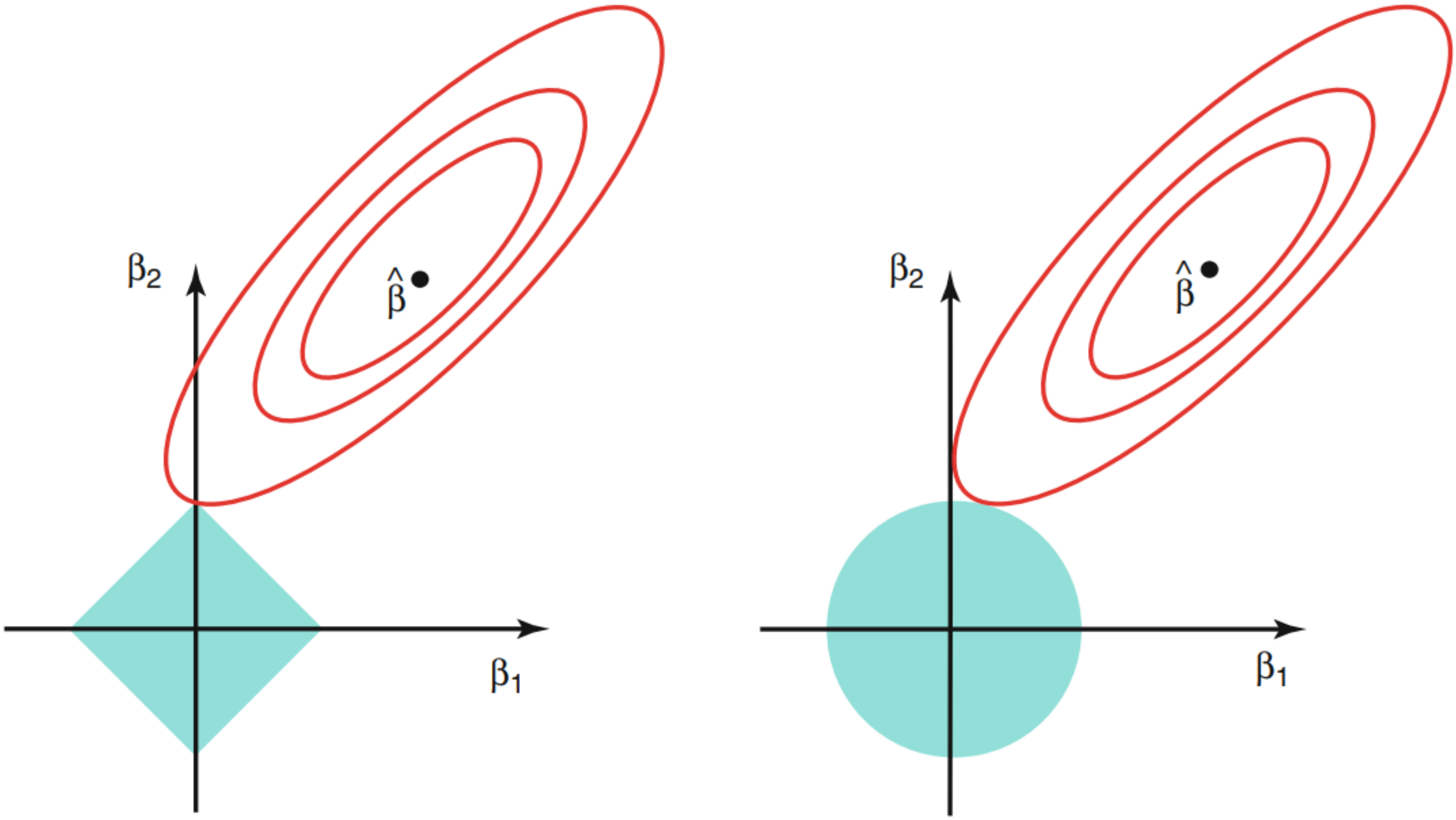


Regressão LASSO - resultados

```
y_lasso <- predict(lasso, newx = X[-idx,], s = cv_lasso$lambda.min)

tab$mse[tab$metodo == "lasso"] <- mean((y[-idx] - y_lasso)^2)

tab
```

```
## # A tibble: 4 x 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lm      11205.
## 2 ridge   14606.
## 3 lasso   11181.
## 4 elastic    NA
```

LASSO e Ridge



Fonte: livro An Introduction to Statistical Learning with Applications in R.

Elastic-net

O elastic-net é uma penalização do tipo

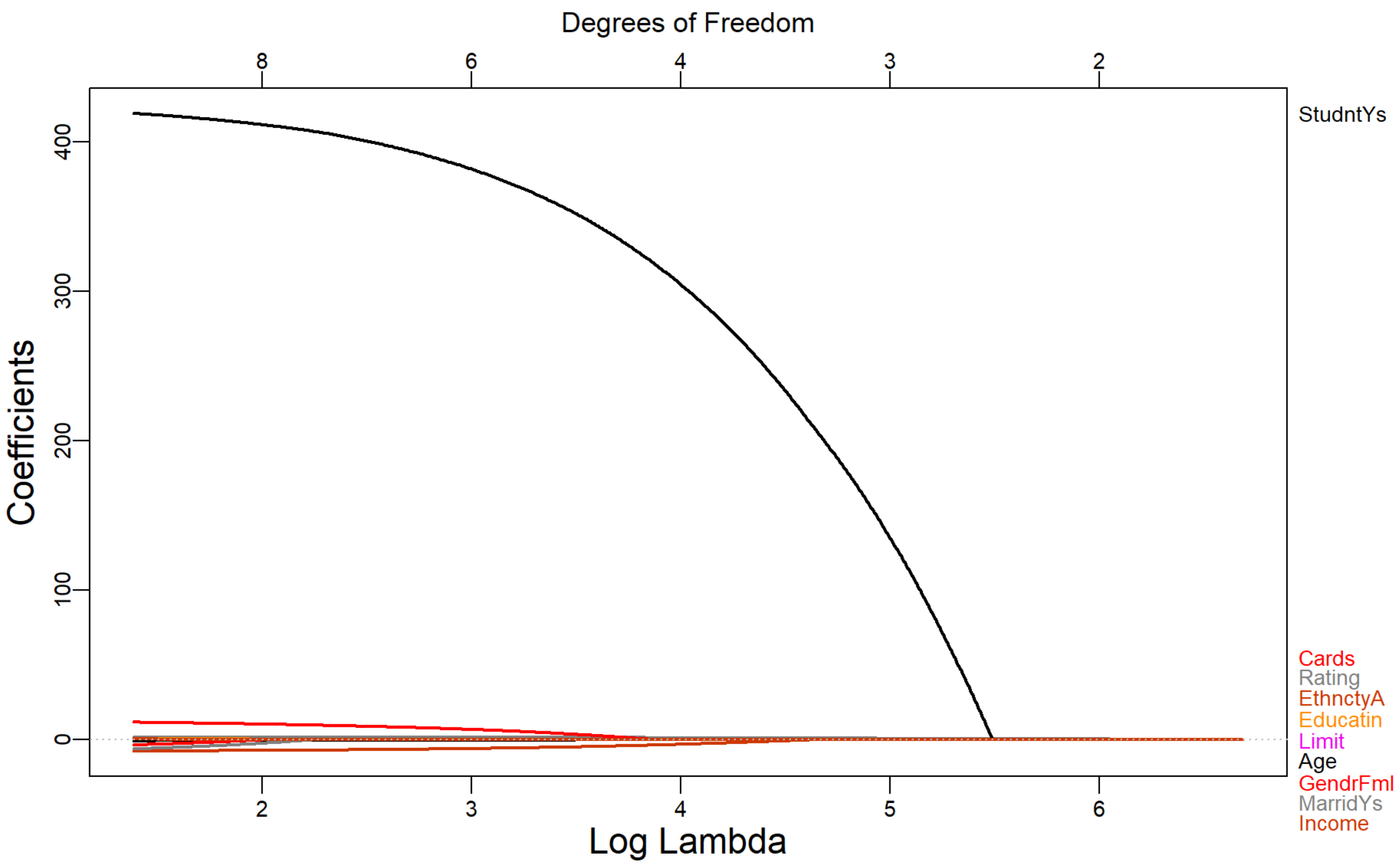
$$\text{RSS} + \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2 \right).$$

Qual a relação desta penalização com o ridge e LASSO?

O que acontece se $\alpha = 0$? E se $\alpha = 1$?

Elastic-net

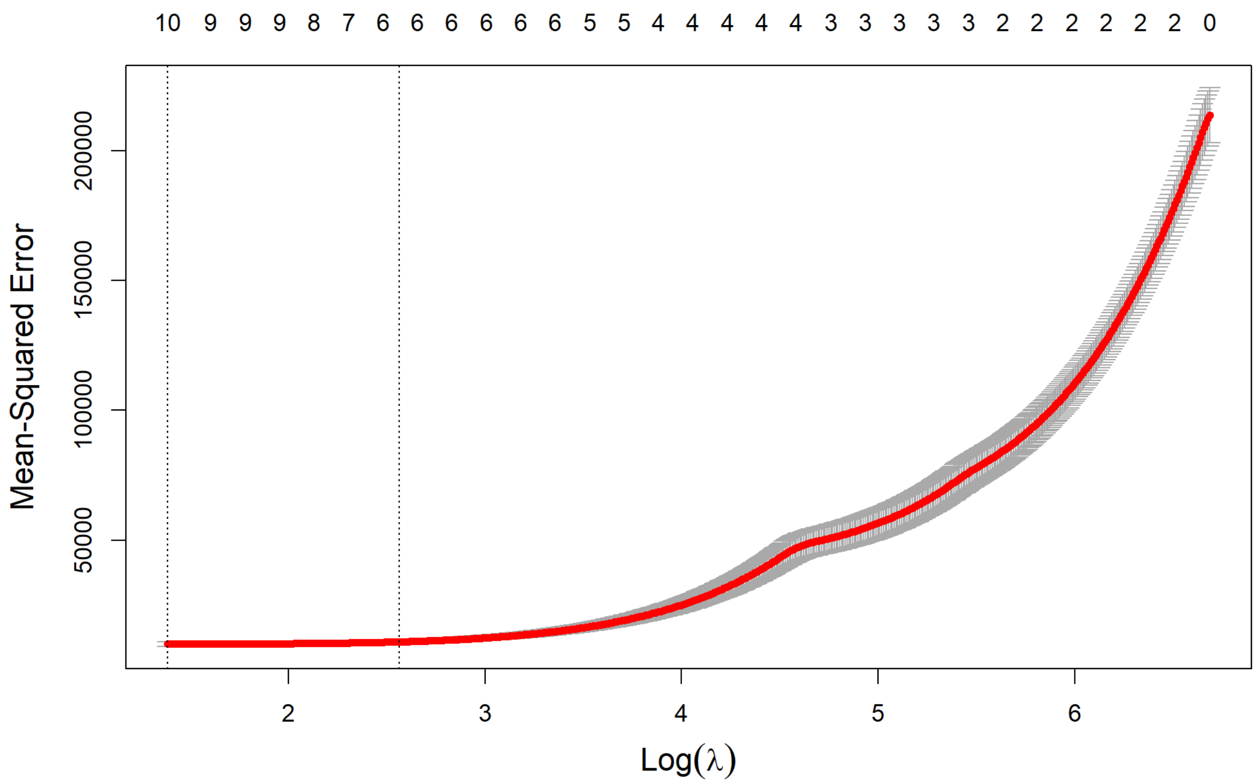
```
elastic <- glmnet(X[idx,], y[idx], alpha = 0.5, nlambda = 1000)
plot_glmnet(elastic, lwd = 2, cex.lab = 1.3, xvar = "lambda")
```



Elastic-net - escolha de λ

Vamos determinar λ através da validação cruzada.

```
cv_elastic <- cv.glmnet(X[idx,], y[idx], alpha = 0.5,
                        lambda = elastic$lambda)
plot(cv_elastic, cex.lab = 1.3)
```

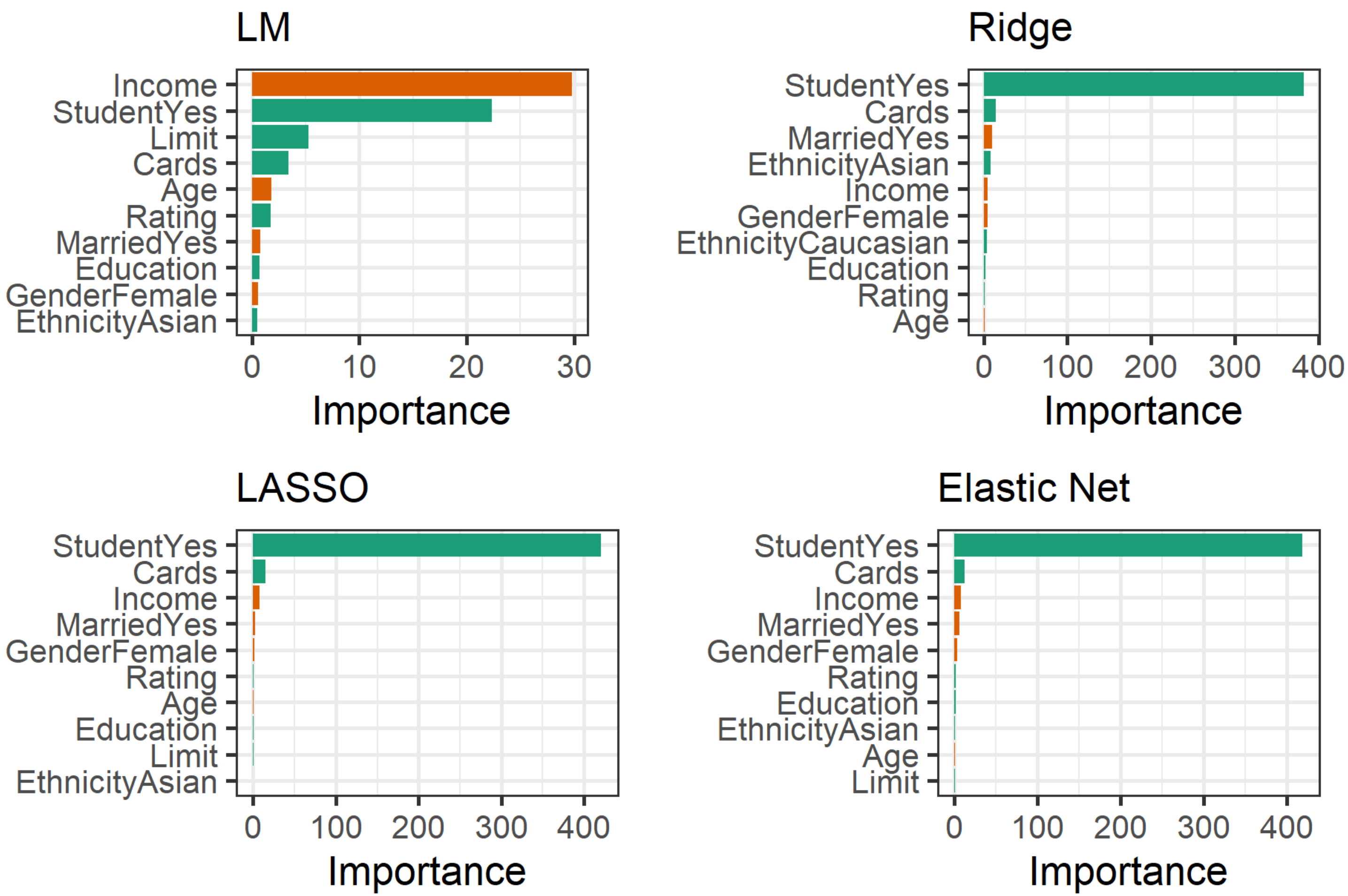


Elastic-net - resultados

```r
y_elastic <- predict(elastic, newx = X[-idx,],
                     s = cv_elastic$lambda.min)
tab$mse[tab$metodo == "elastic"] <- mean((y[-idx] - y_elastic)^2)
tab
```

```
## # A tibble: 4 x 2
##   metodo     mse
##   <chr>    <dbl>
## 1 lm      11205.
## 2 ridge   14606.
## 3 lasso   11181.
## 4 elastic 11201.
```

Importância de variáveis



Pratica R

Problemas de classificação

Problemas de classificação

- Situações em que o objetivo é assinalar uma classe à uma observação.
- Dados `Default` [1]:
  - Informações sobre 1000 clientes;
  - **default**: indica se o cliente apresentou *default*;
  - **student**: indica se o cliente é estudante;
  - **balance**: saldo médio mensal no cartão de crédito;
  - **income**: renda do cliente;
- Objetivo: predizer quais clientes apresentarão default no cartão de crédito.

[1] Fonte: dados no pacote `ISLR`, do livro An Introduction to Statistical Learning with Applications in R.

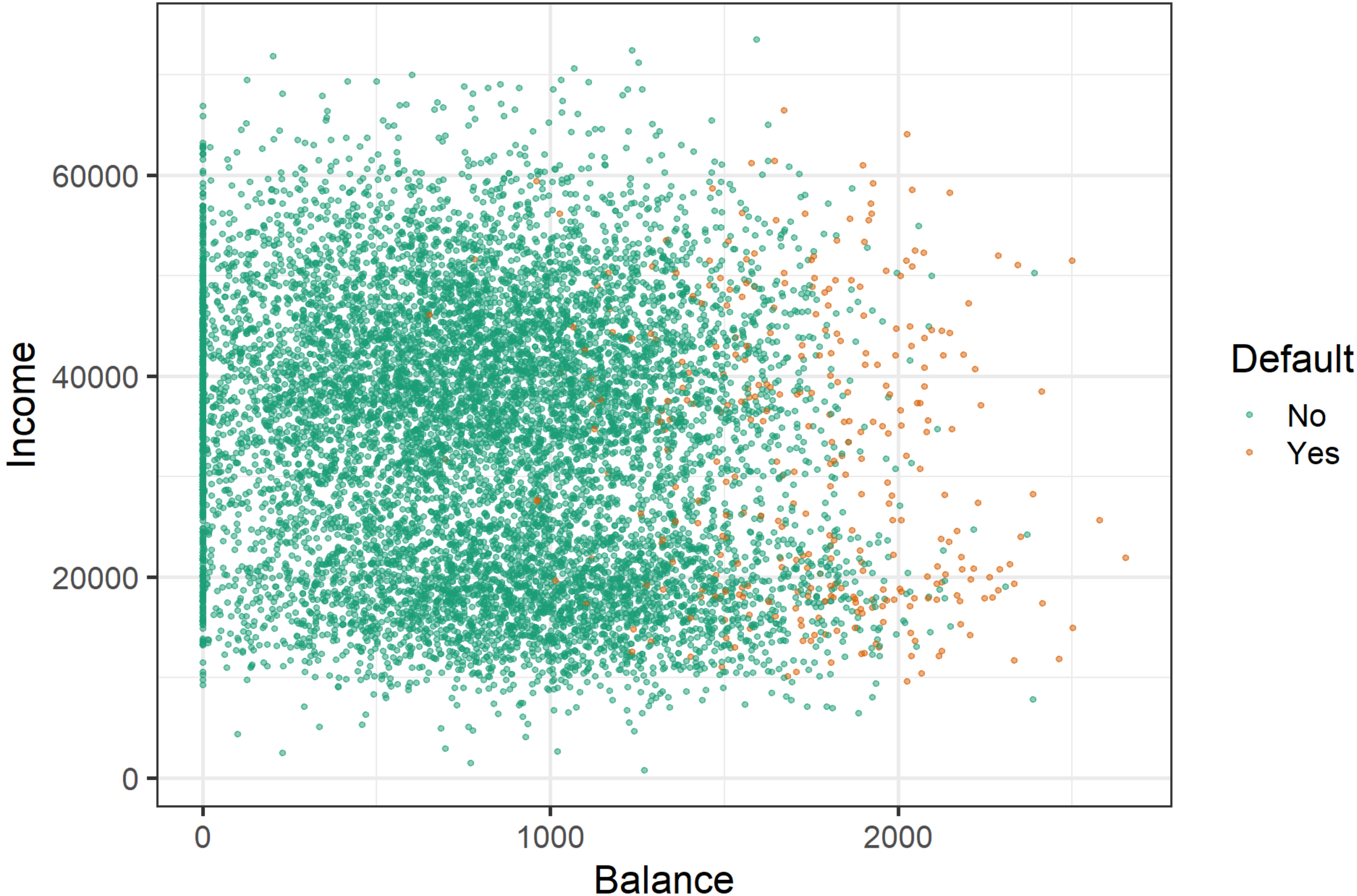
Dados `Default`

	default	student	balance	income
1	No	No	729.53	44361.63
2	No	Yes	817.18	12106.13
3	No	No	1073.55	31767.14
4	No	No	529.25	35704.49
5	No	No	785.66	38463.5
6	No	Yes	919.59	7491.56
7	No	No	825.51	24905.23
8	No	Yes	808.67	17600.45
9	No	No	1161.06	37468.53
10	No	No	0	29275.27

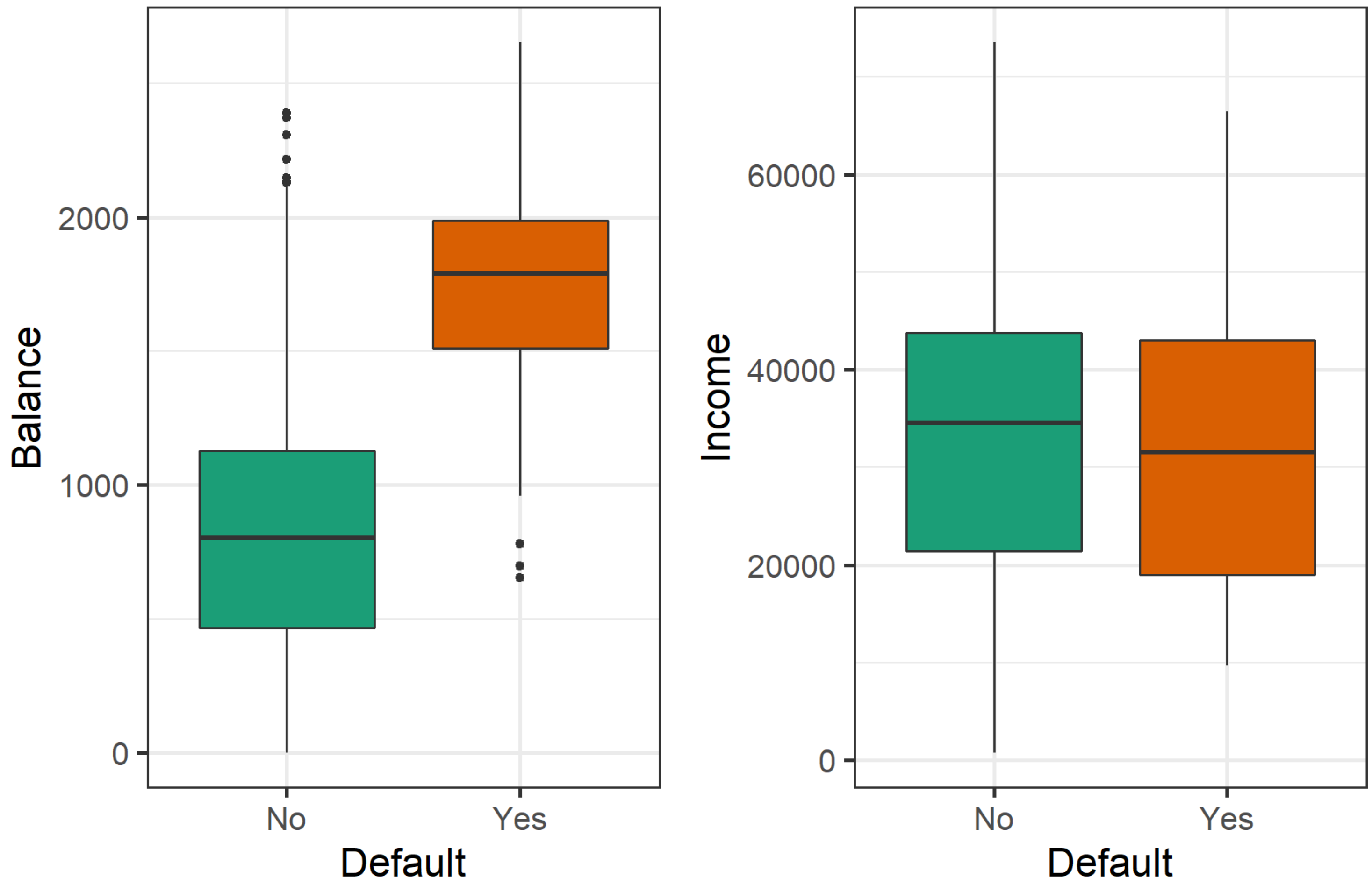
Showing 1 to 10 of 10,000 entries

Previous 1 2 3 4 5 … 1000 Next

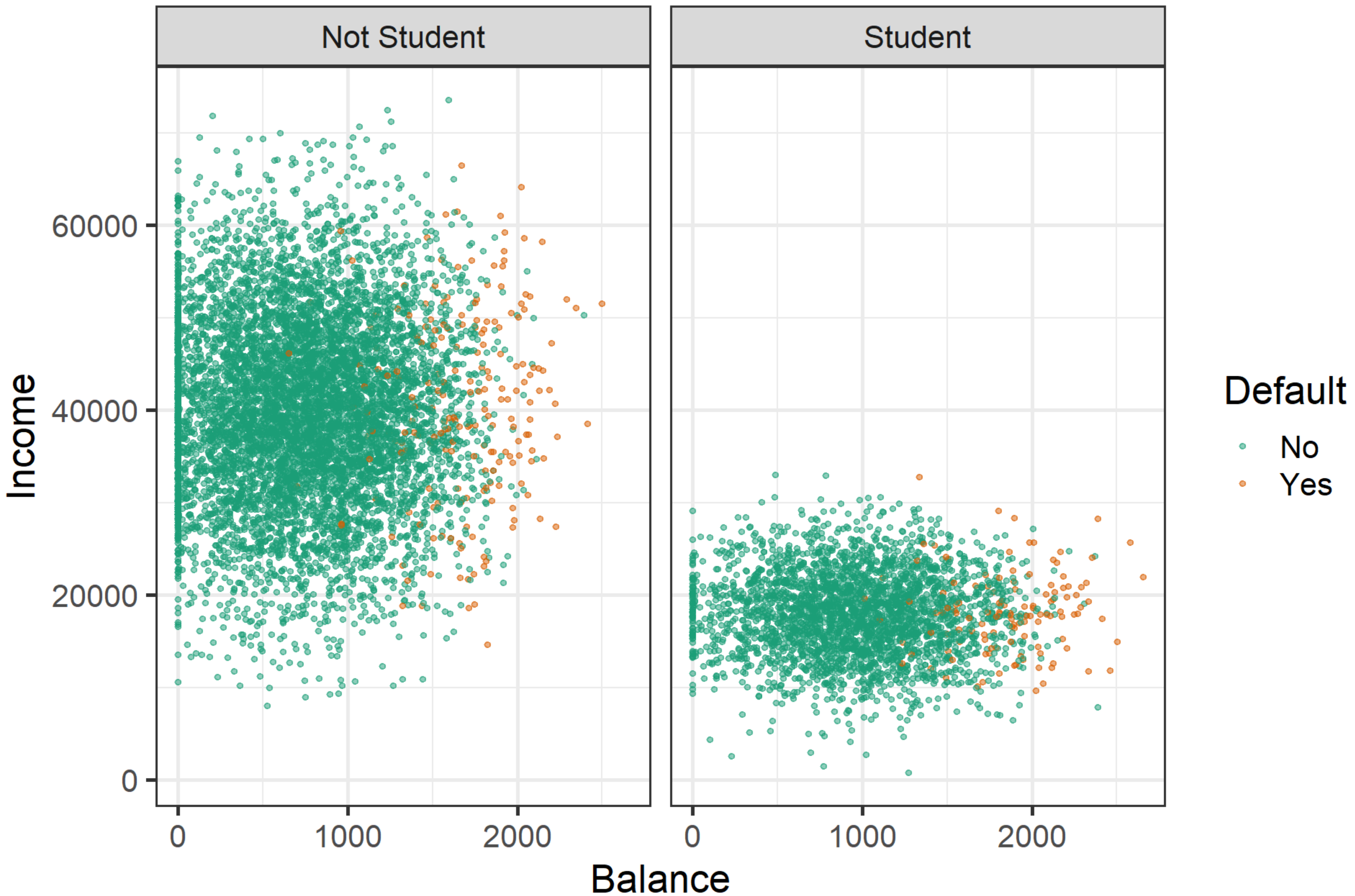
Análise exploratória



Análise exploratória

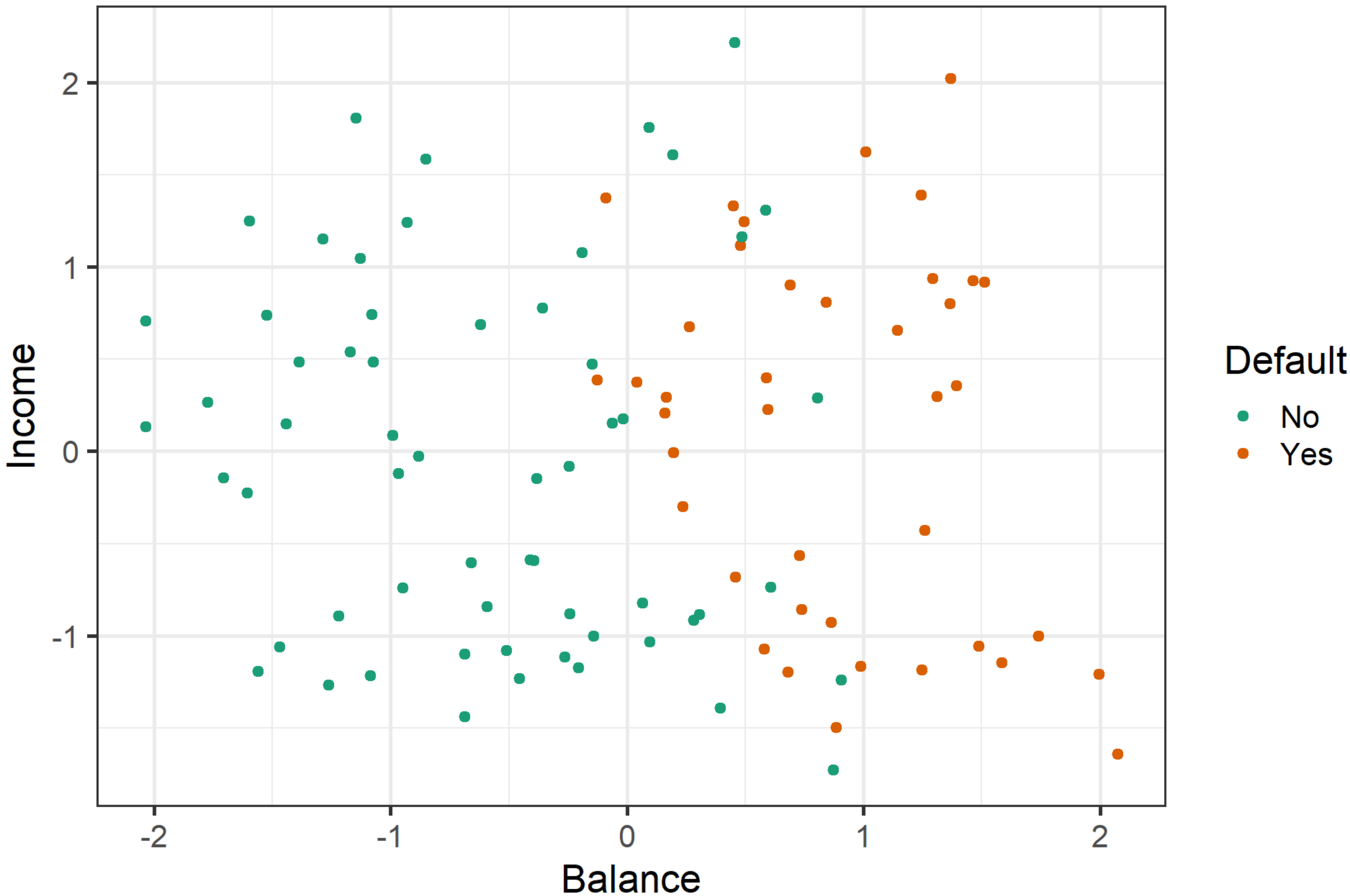


Análise exploratória



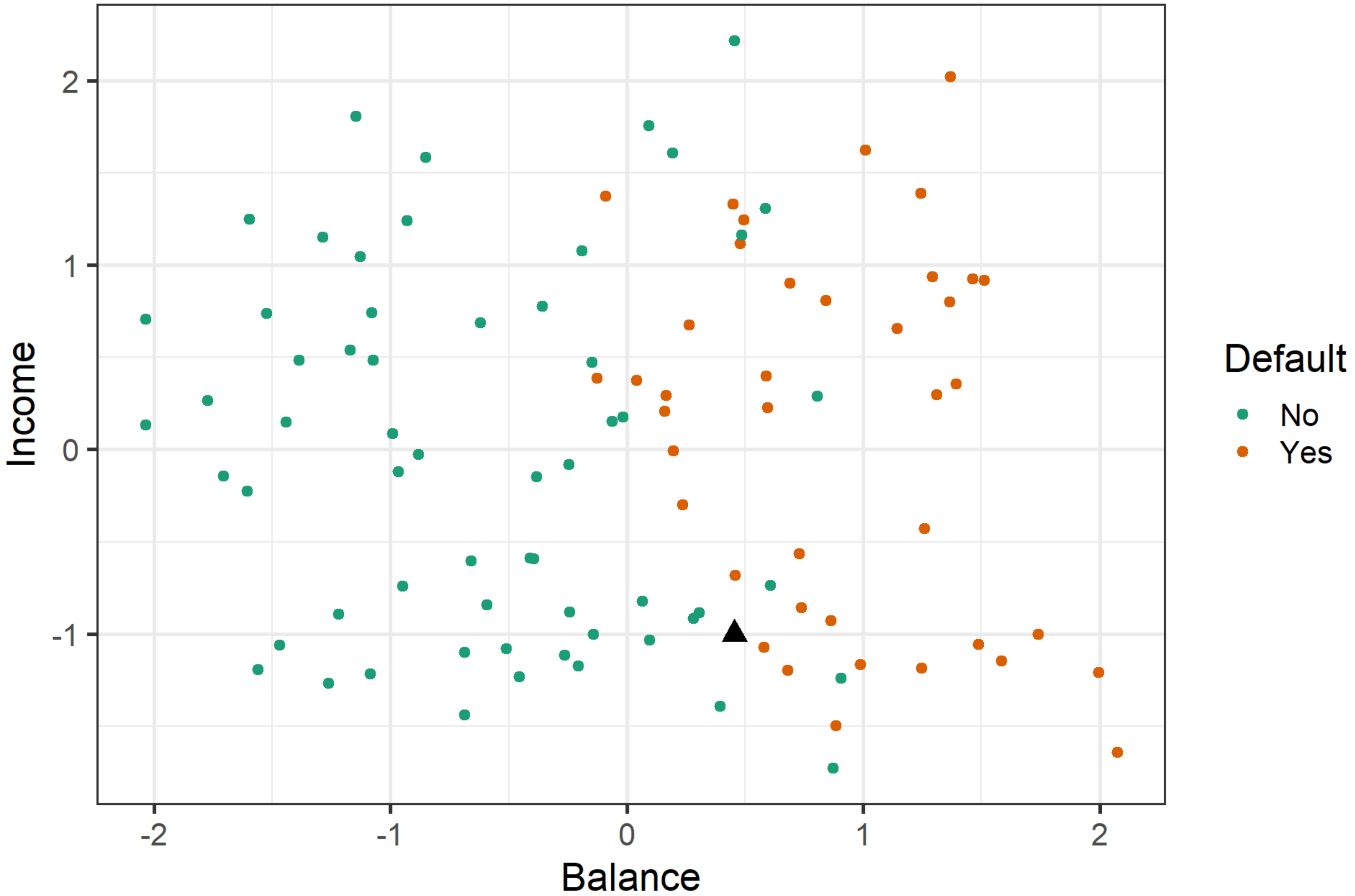
KNN para classificação

KNN para classificação



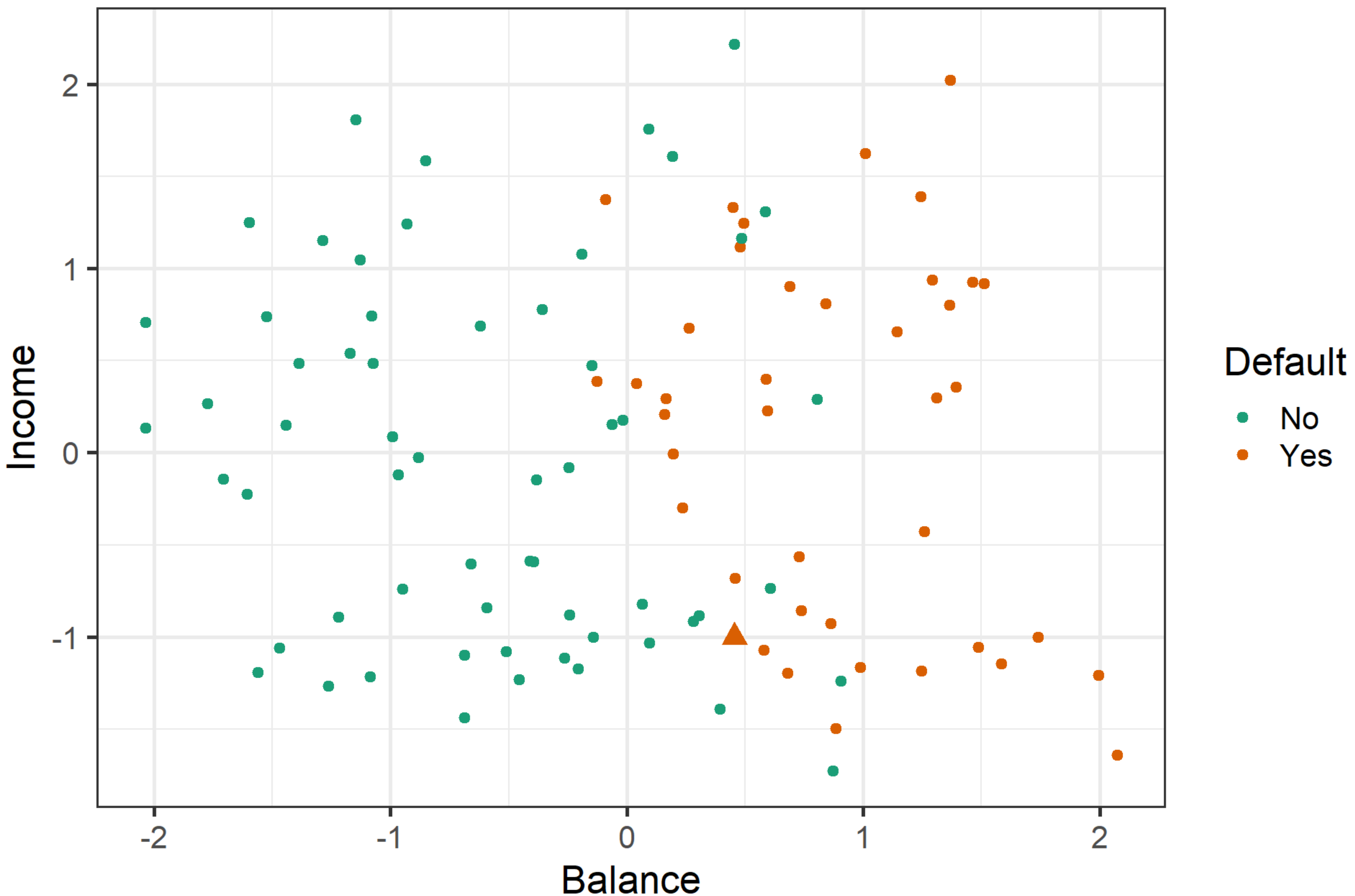
KNN para classificação

Qual seria a previsão para o ponto preto no gráfico?



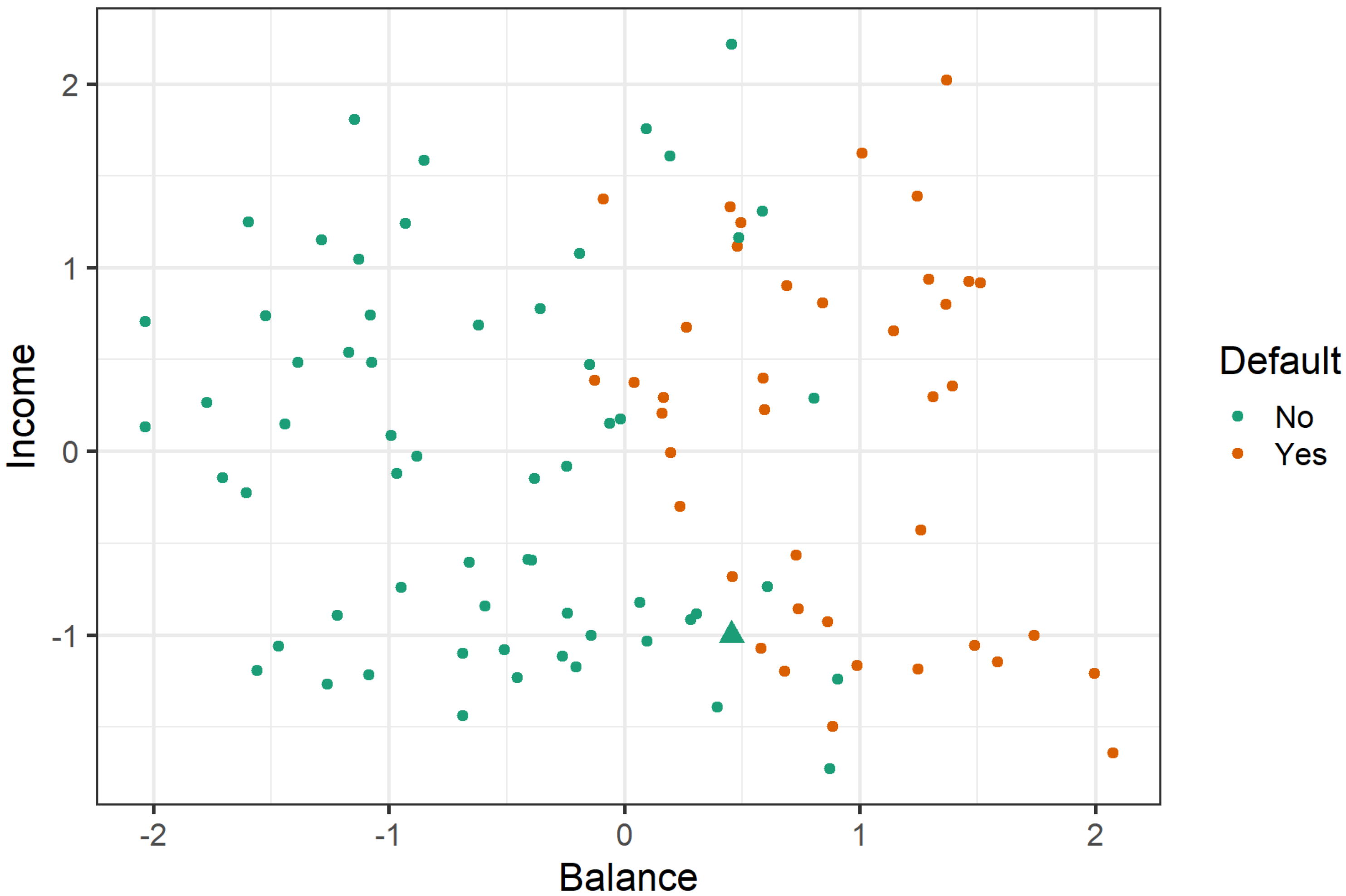
KNN para classificação

Considerando $k = 1$.



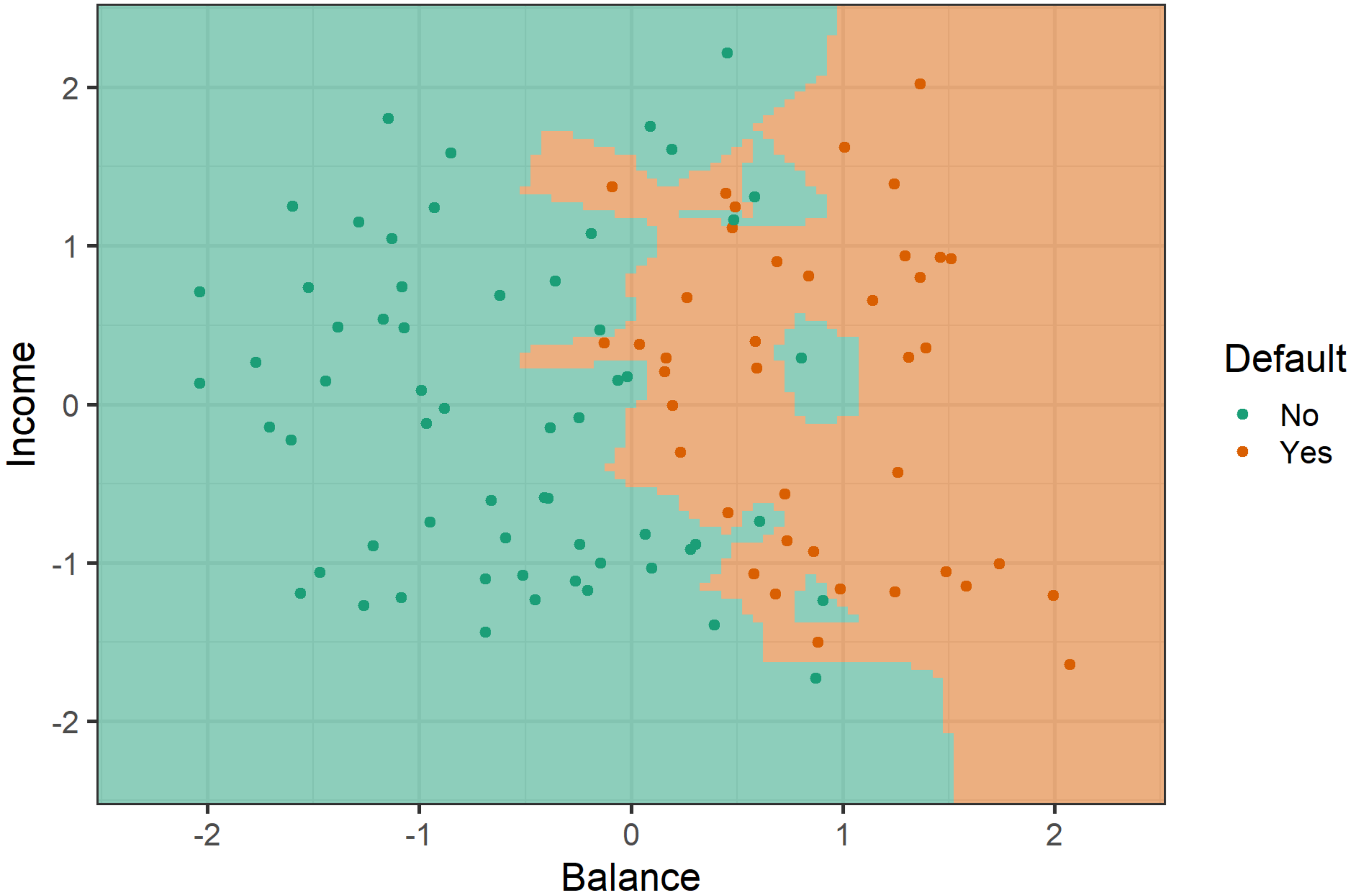
KNN para classificação

Considerando $k = 3$.



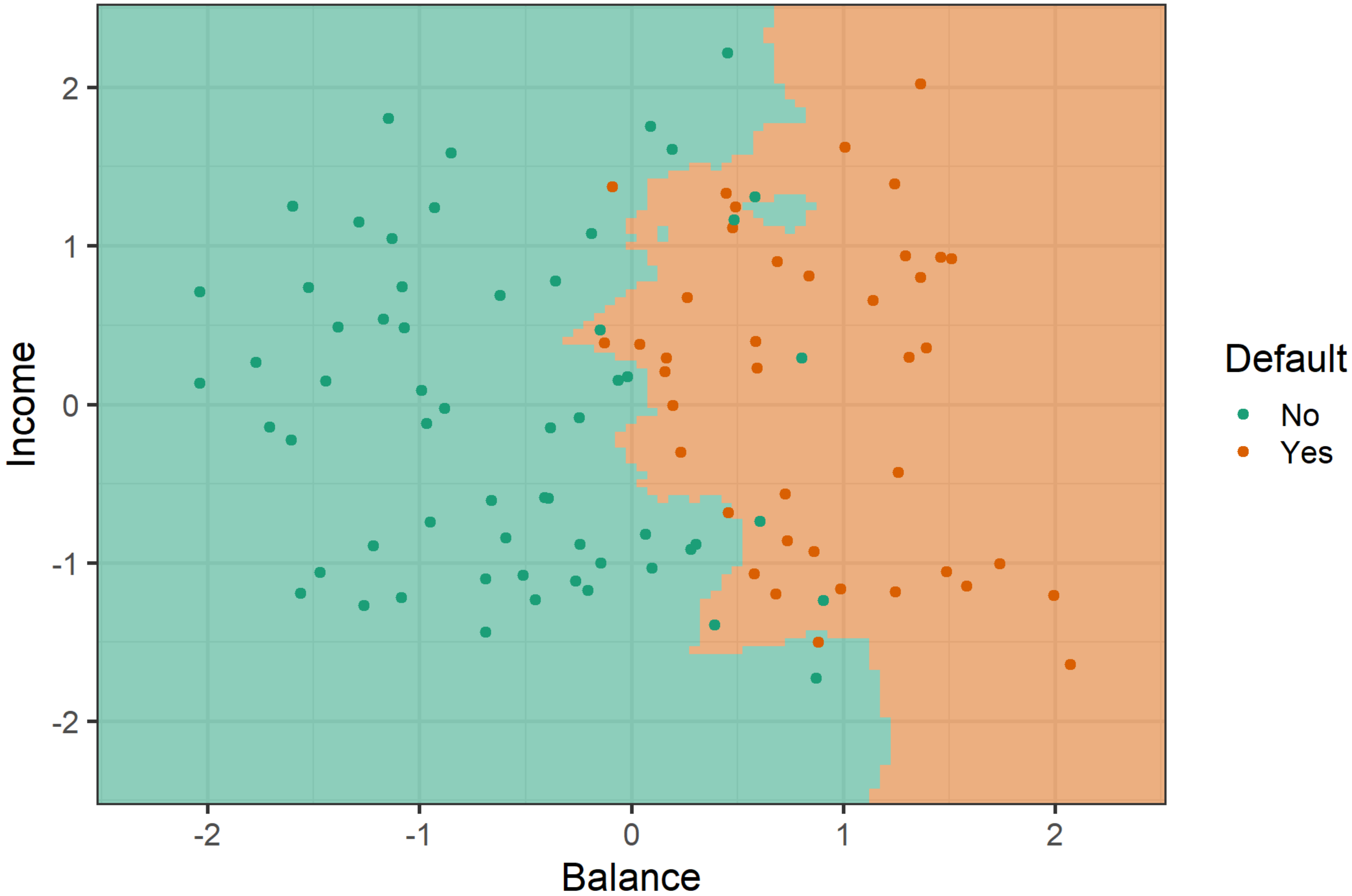
KNN para classificação

Considerando $k = 1$.



KNN para classificação

Considerando $k = 3$.



Regressão logística

Classificação

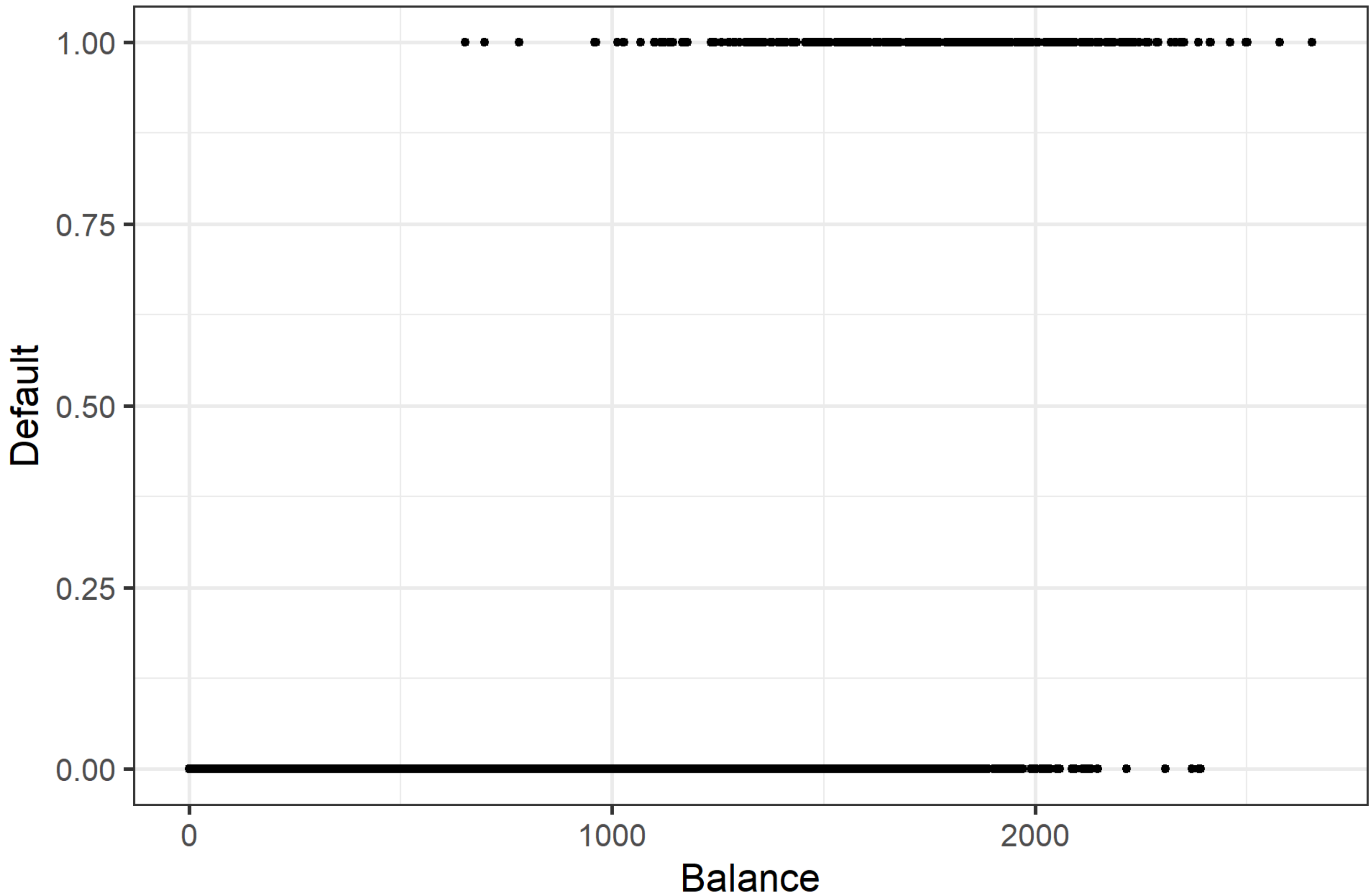
- A variável `default` assume dois possíveis valores: `Yes` e `No`.
- Ao invés de modelar diretamenteo a variável resposta Y, vamos modelar a *probabilidade* de Y pertencer a uma categoria em particular.
- Por exemplo, a probabilidade de `default = Yes` dado a variável `balance` pode ser escrita como

$$p(\text{balance}) = P(\text{default} = Yes|\text{balance}).$$

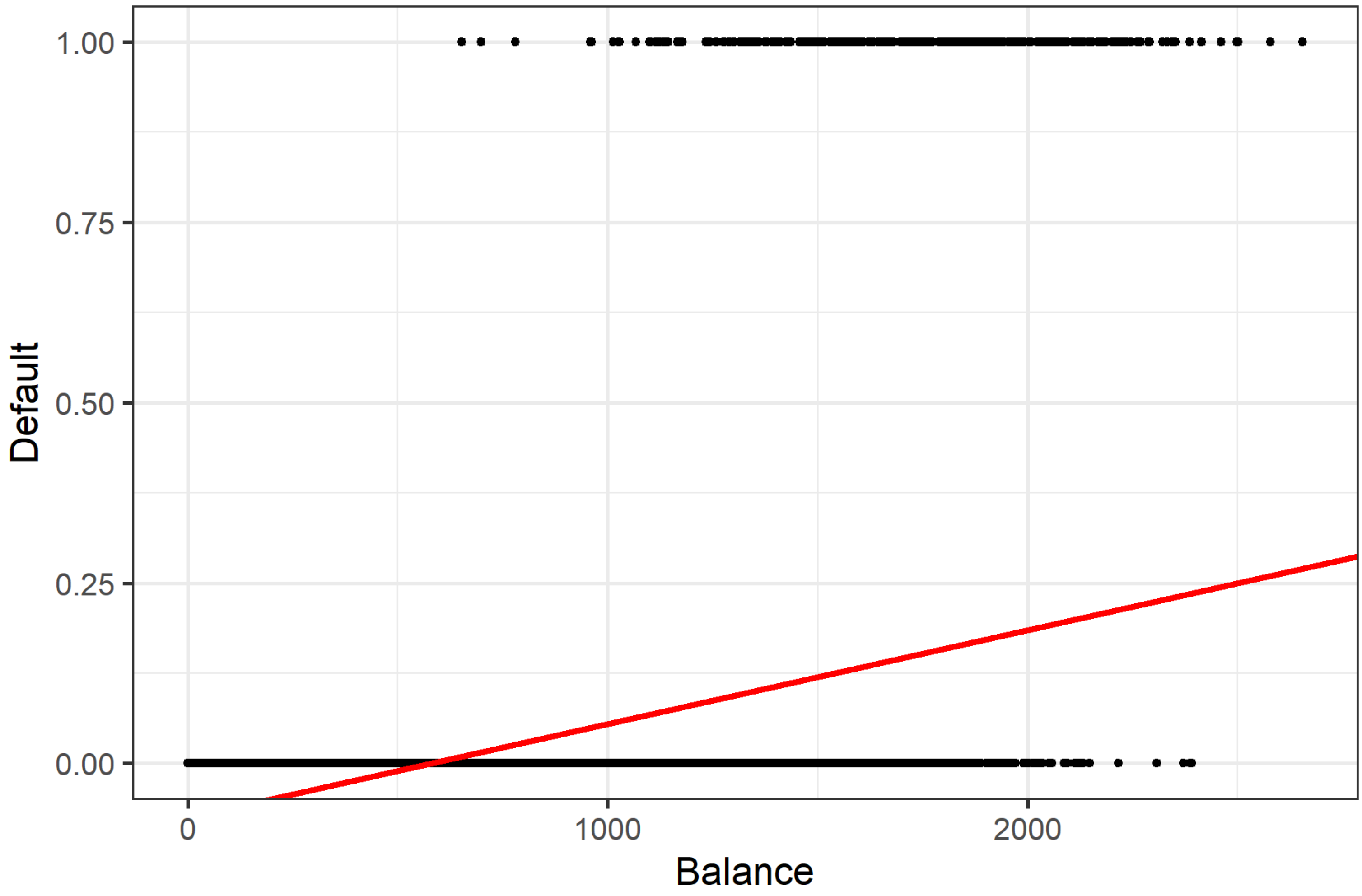
- Poderíamos, por exemplo modelar $p(\text{balance})$ por

$$p(\text{balance}) = \beta_0 + \beta_1 \times \text{balance}.$$

Classificação



Porque não usar regressão linear?



Problema?

Alternativa: modelar uma função da chance

$$\text{chance} = \frac{p(X)}{1 - p(X)}.$$

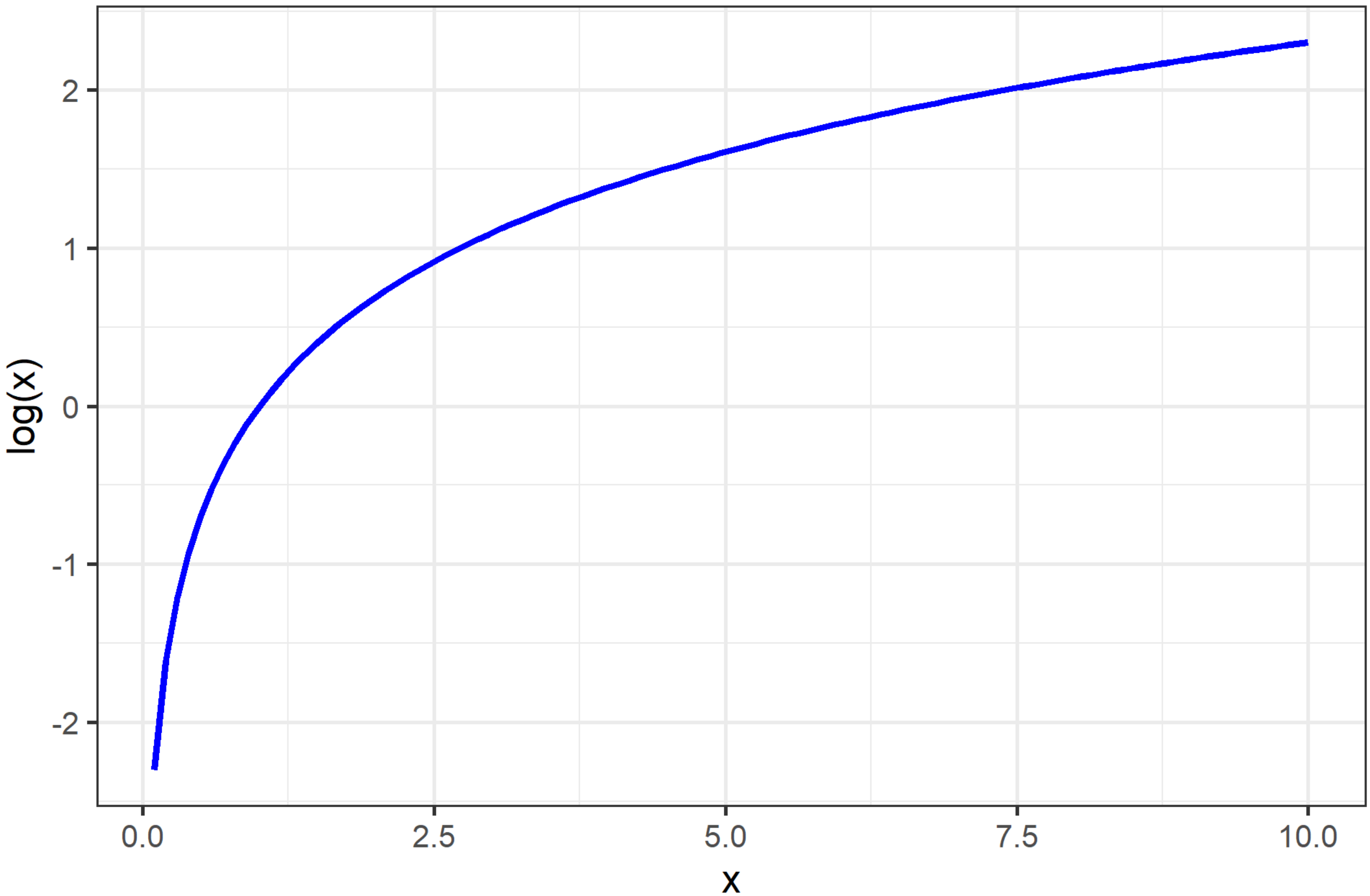
Probabilidade	Chance
0.90	90:10 ou 9
0.75	75:25 ou 3
0.50	50:50 ou 1
0.20	20:80 ou 0.25
0.10	10:90 ou 0.11
0.01	1:99 ou 0.01

Alternativa: modelar uma função da chance

$$\text{chance} = \frac{p(X)}{1 - p(X)}.$$

Probabilidade	Chance	Log da chance
0.90	90:10 ou 9	2.197
0.75	75:25 ou 3	1.099
0.50	50:50 ou 1	0.000
0.20	20:80 ou 0.25	-1.386
0.10	10:90 ou 0.11	-2.197
0.01	1:99 ou 0.01	-4.595

Alternativa: modelar uma função da chance



Log da chance

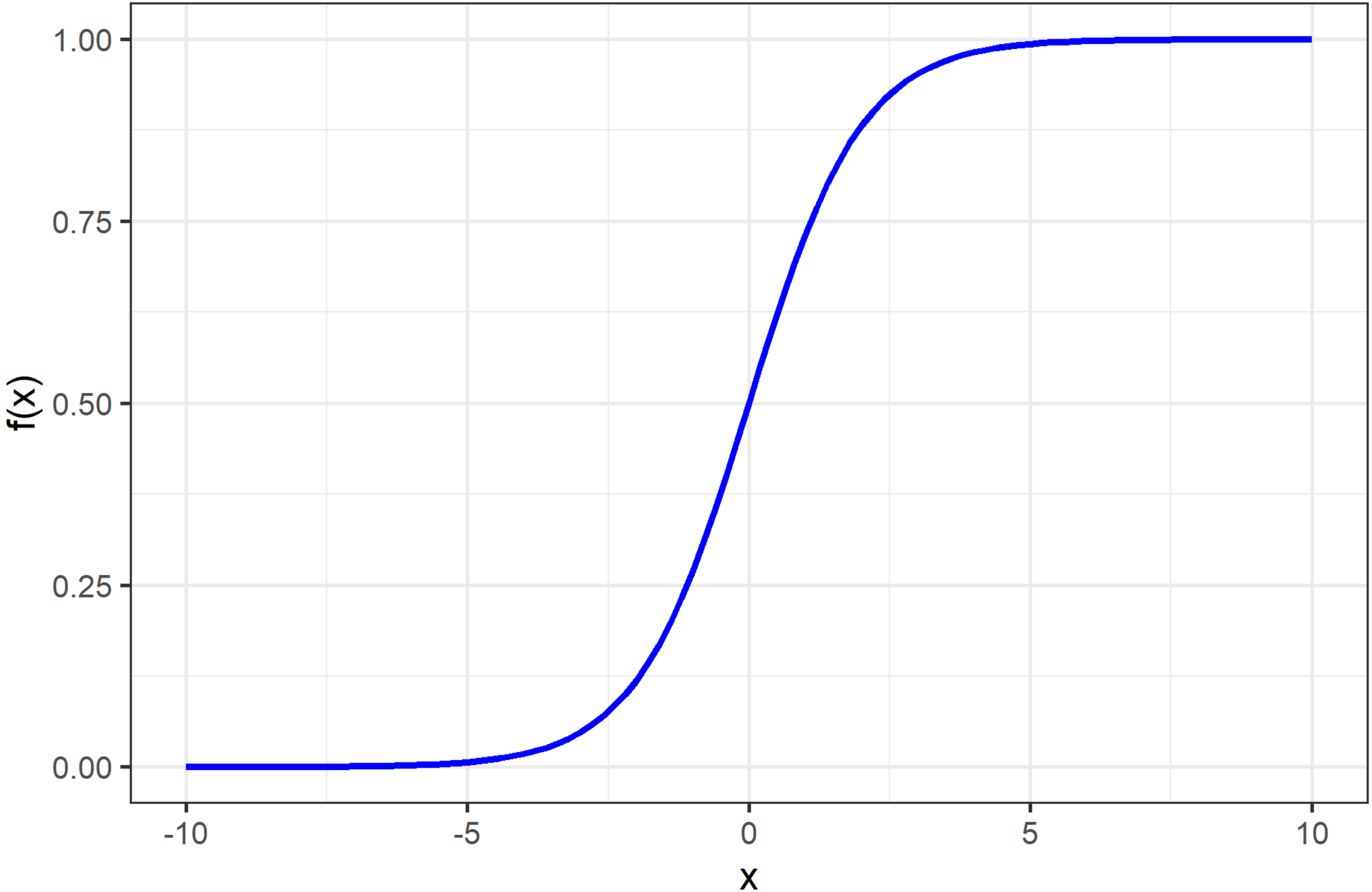
$$\log(\text{chance}) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

Após algumas manipulções algébricas para isolar $p(X)$, temos que

$$p(X) = \frac{\exp\{\beta_0 + \beta_1 X\}}{1 + \exp\{\beta_0 + \beta_1 X\}} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 X)\}}.$$

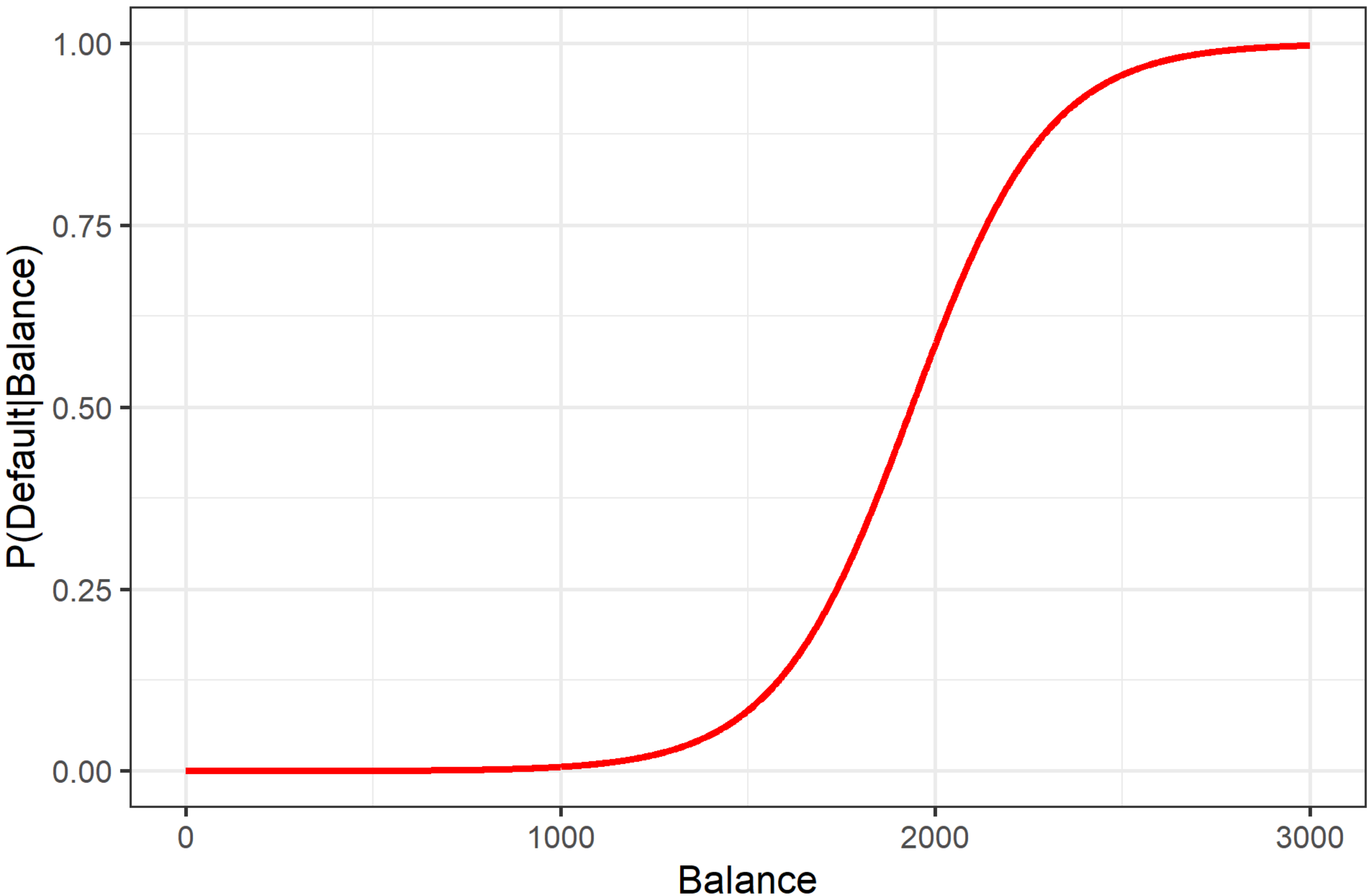
Logística

$$f(x) = \frac{\exp\{x\}}{1 + \exp\{x\}}$$



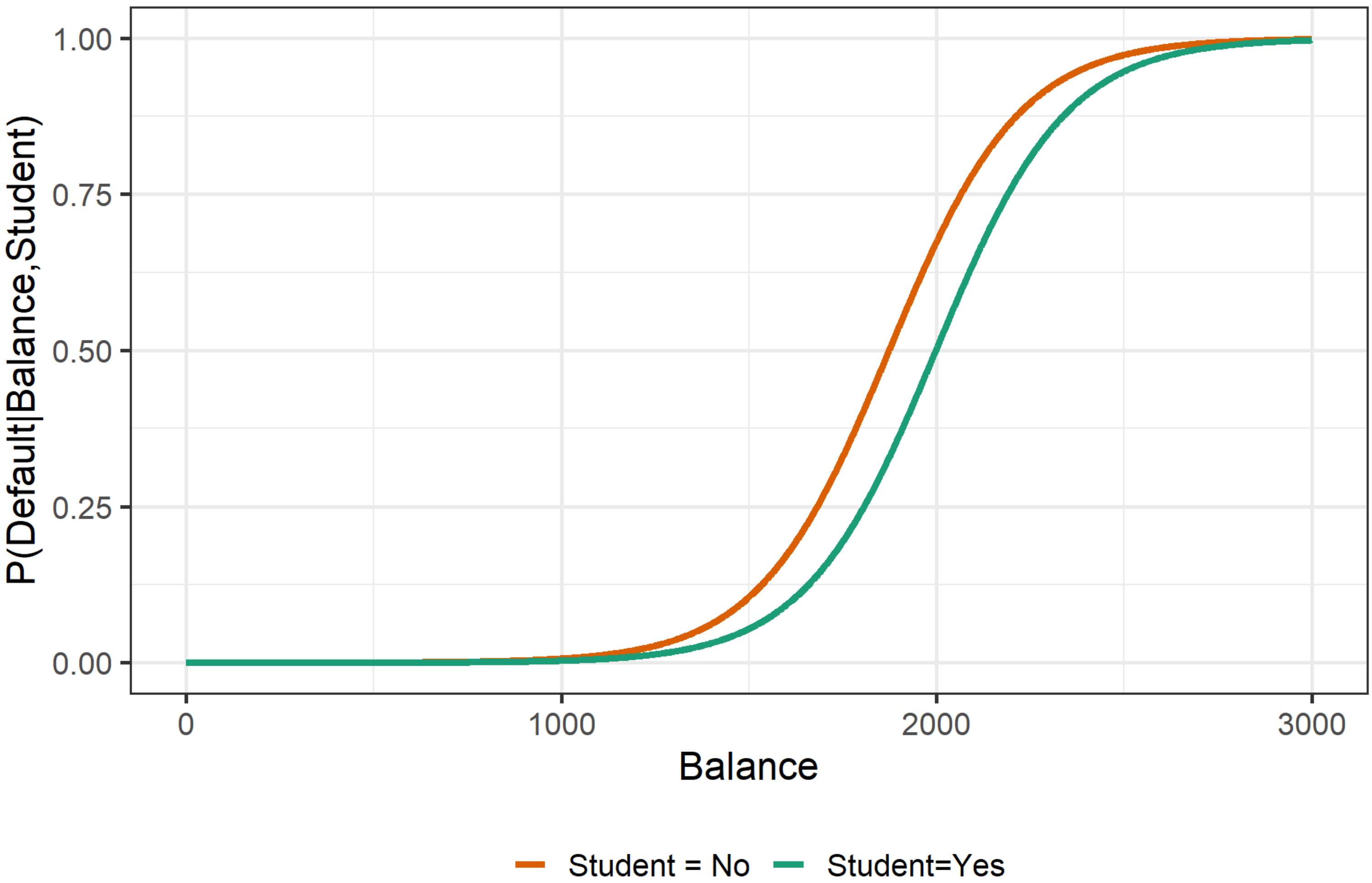
Logística

$$\log\left(\frac{p(x)}{1-p(x)}\right) = -10.6513 + 0.0055x.$$



Logística

$$\log\left(\frac{p(x)}{1-p(x)}\right) = -10.7495 + 0.0057 \times balance - 0.7149 \times student.$$



Como obter as estimativas para β_j?

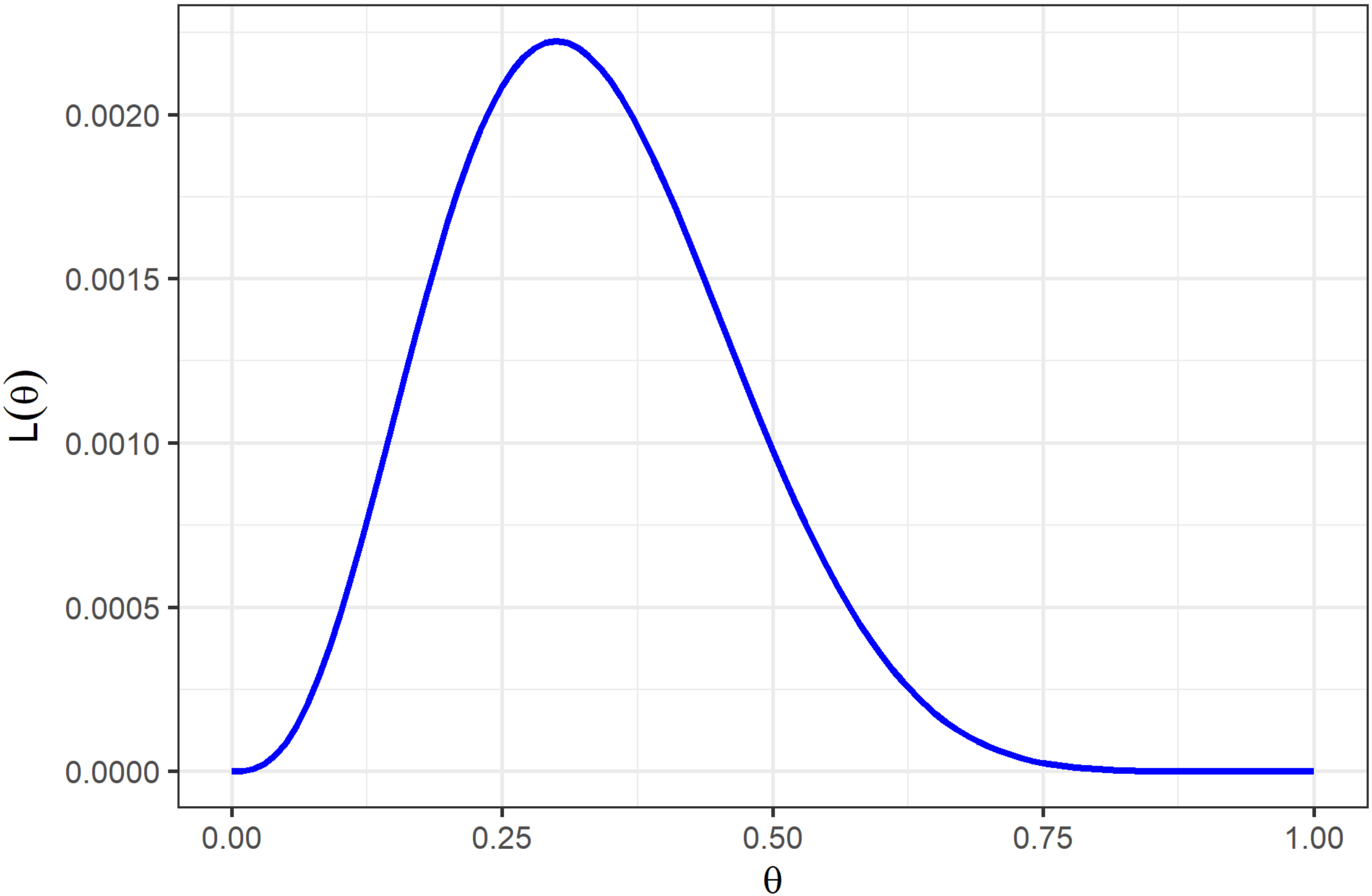
Vamos considerar uma situação em que todas observações apresentam a mesma probabilidade θ de apresentar *default* e foram observados d *defaults* numa amostra de tamanho n. Assim,

$$
\begin{aligned}
L_y(\theta) &= P(Y_1 = y_1, \ldots, Y_n = y_n|\theta) \\
&= \prod_{i=1}^{n} P(Y_i = y_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\
&= \theta^d (1-\theta)^{n-d}.
\end{aligned}
$$

Exemplo

Tome n=10 e d=3. Então

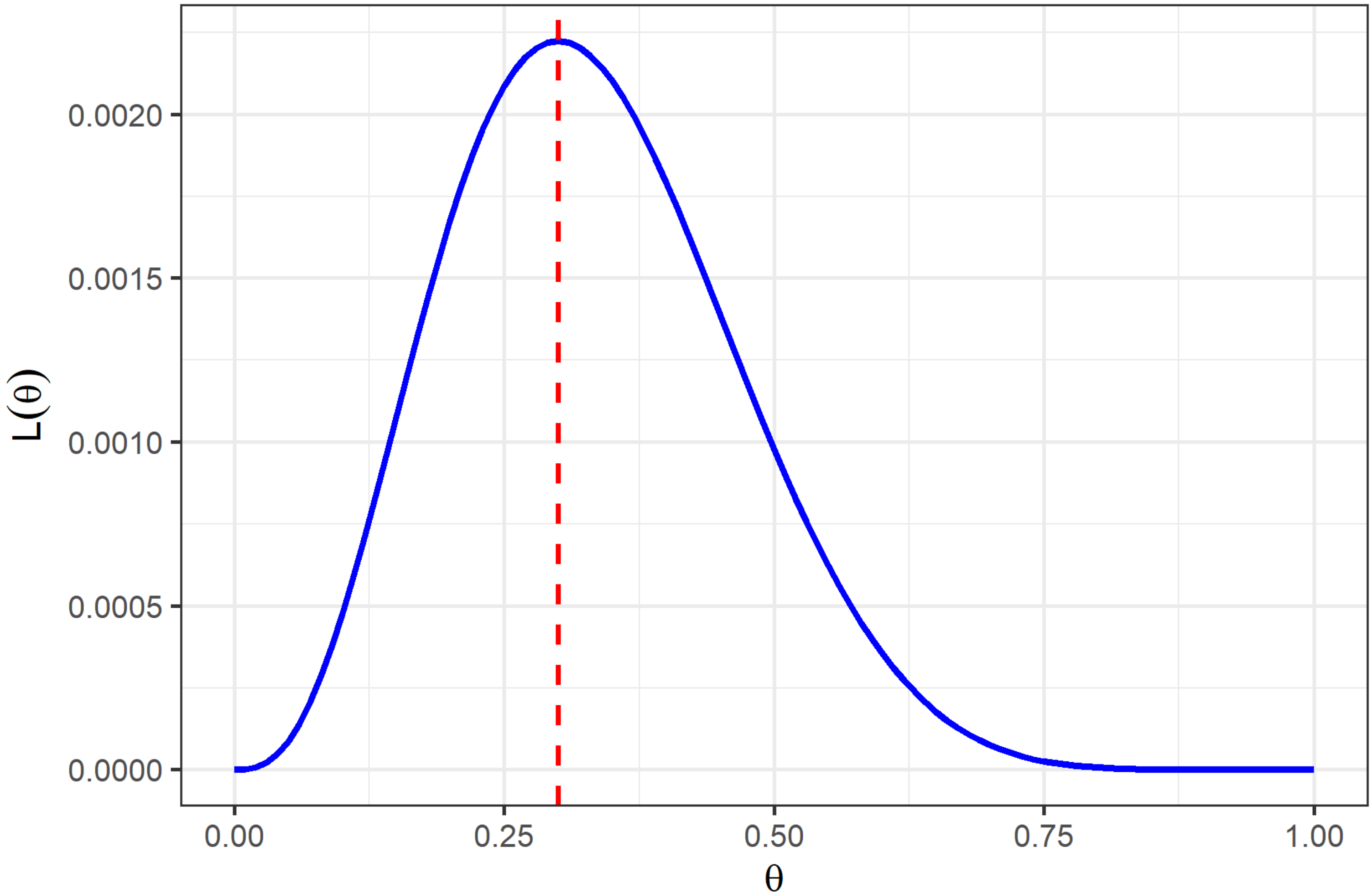
$$L_y(\theta) = \theta^3(1-\theta)^{10-3}.$$



Exemplo

Tome n=10 e d=3. Então

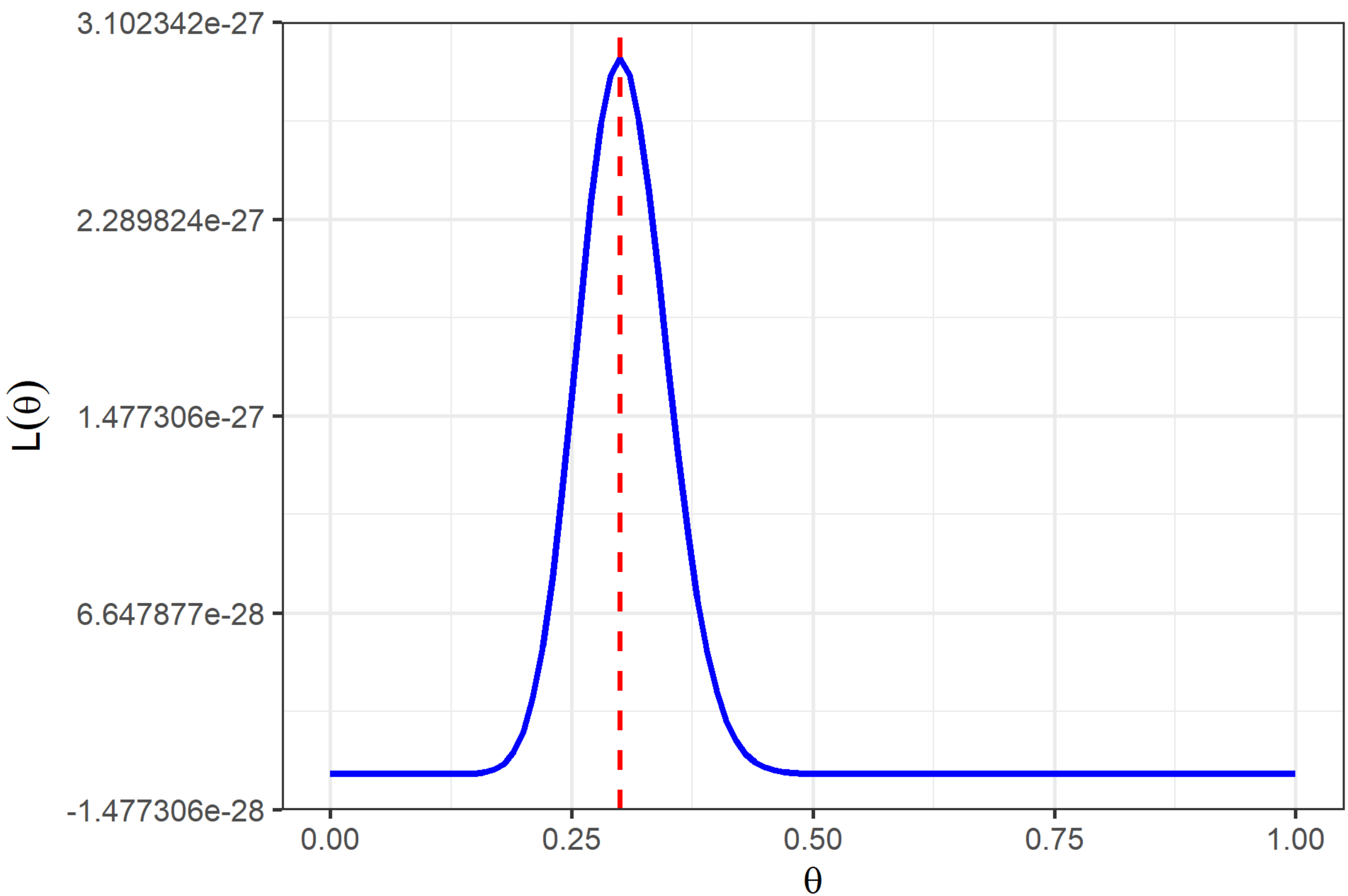
$$L_y(\theta) = \theta^3(1-\theta)^{10-3}.$$



Exemplo

Tome n=100 e d=30. Então

$$L_y(\theta) = \theta^{30}(1-\theta)^{100-30}.$$



Como obter as estimativas para β_j?

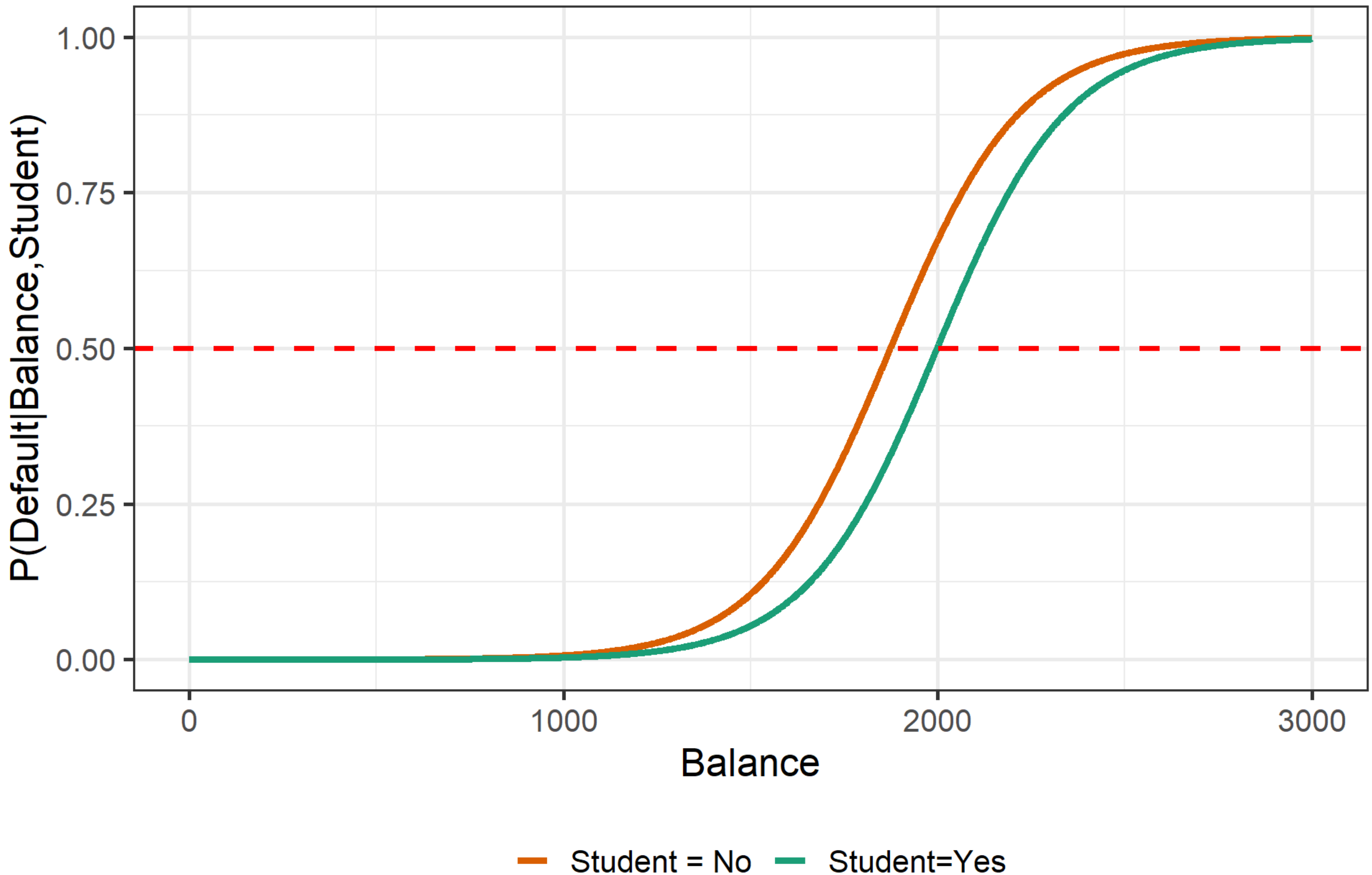
Com a função de verossimilhança.

$$
\begin{aligned}
L_{\mathbf{x},y}(\theta) &= P(Y_1 = y_1, \ldots, Y_n = y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \beta) \\
&= \prod_{i=1}^{n} P(Y_i = y_i | \mathbf{x}_i, \theta) \\
&= \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i} [1 - p(\mathbf{x}_i)]^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{\exp\{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_1 x_{p,i}\}}{1 + \exp\{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_1 x_{p,i}\}} \right)^{y_i} \left(1 - \frac{\exp\{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_1 x_{p,i}\}}{1 + \exp\{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_1 x_{p,i}\}} \right)^{1-y_i}.
\end{aligned}
$$

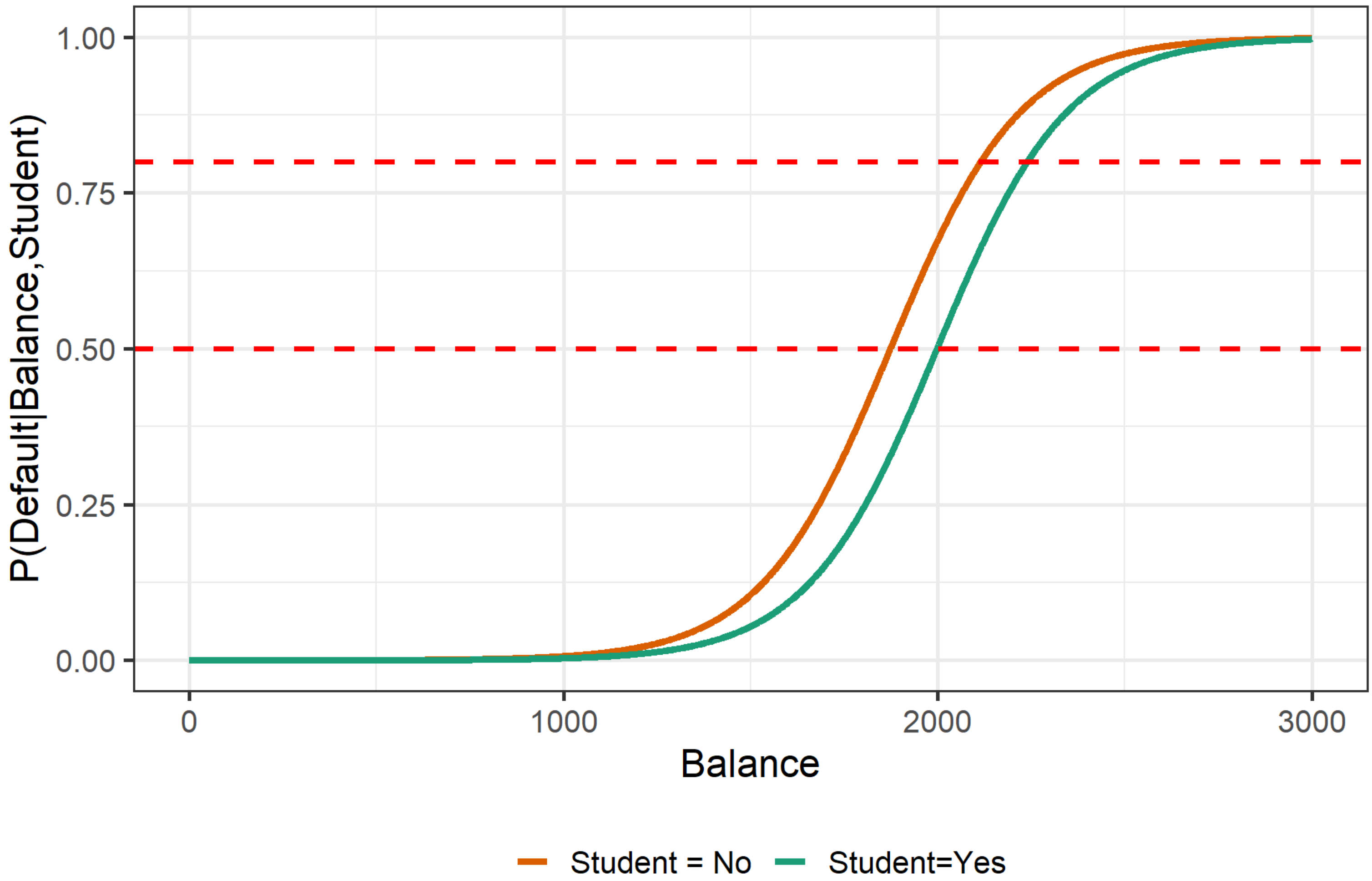
Modelo estimado

```
##
## Call:
## glm(formula = default ~ balance + student, family = binomial,
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4578  -0.1422  -0.0559  -0.0203   3.7435
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
## balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
## studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.7  on 9997  degrees of freedom
## AIC: 1577.7
##
## Number of Fisher Scoring iterations: 8
```

Como classificar?



Como classificar?



Matriz de confusão

Para corte = 0.5.

```
##        Observado
## Predito   No  Yes
##      No 9628  228
##      Yes  39  105
```

Para corte = 0.8.

```
##        Observado
## Predito   No  Yes
##      No 9663  303
##      Yes   4   30
```

Métricas

	Observado	
Classificado	**No**	**Yes**
No	a	b
Yes	c	d

Erro de classificação total: $\frac{b+c}{n} = 1 - \frac{a+d}{n}$;

Verdadeiro positivo (sensibilidade ou recall: $\frac{d}{b+d}$;

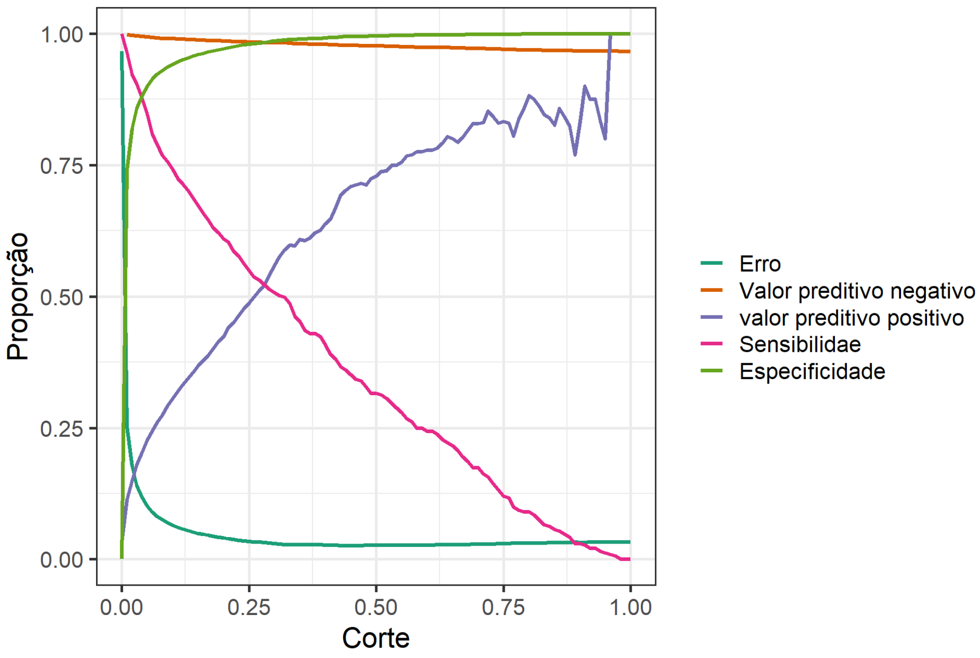
Verdadeiro negativo (especificidade): $\frac{a}{a+c}$;

Valor preditivo positivo (precision): $\frac{d}{c+d}$;

Valor preditivo negativo: $\frac{a}{a+b}$;

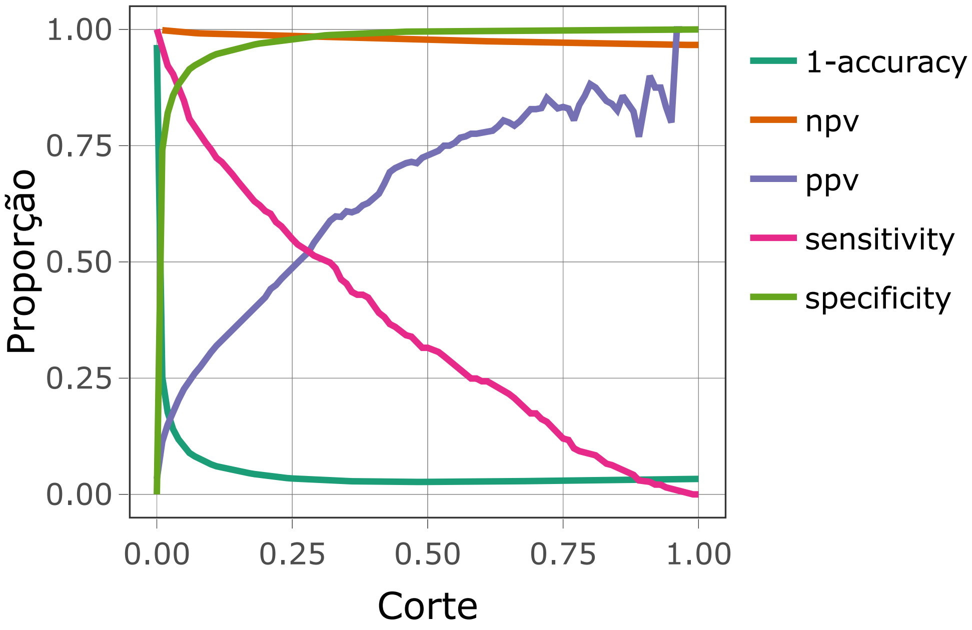
F-score: $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

Medidas

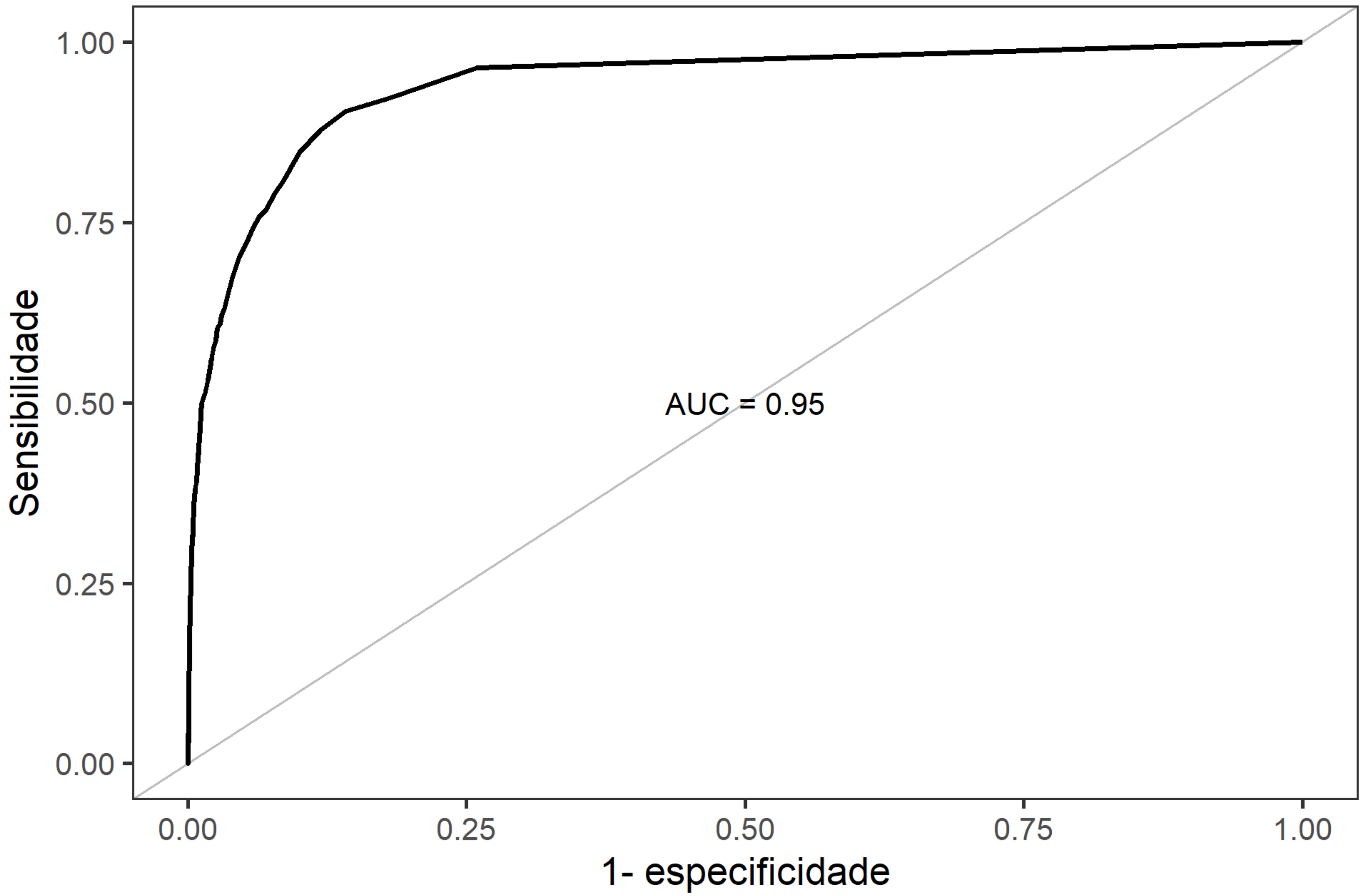


1.00
0.75
0.50
0.25
0.00
Proporção
0.00
0.25
0.50
0.75
1.00
Corte
Erro
Valor preditivo negativo
valor preditivo positivo
Sensibilidae
Especificidade

Medidas


1.00
0.75
0.50
0.25
0.00
Proporção
0.00
0.25
0.50
0.75
1.00
Corte
1-accuracy
npv
ppv
sensitivity
specificity

Curva ROC



1.00
0.75
0.50
0.25
0.00
Sensibilidade
AUC = 0.95
0.00
0.25
0.50
0.75
1.00
1- especificidade

Adicionando fator de perda/ganho na classificação

Considere os seguintes ganhos/perdas a depender da classificação feita por um dado modelo.

	Observado	
Classificado	**No**	**Yes**
No	10	-5
Yes	-20	100

Em um conjunto com 100 observações, obteve-se o seguinte cenário.

	Observado	
Classificado	**No**	**Yes**
No	30	20
Yes	10	40

Então, o lucro esperado será

$$\begin{aligned}
\text{Lucro esperado} &= 10 \times \frac{30}{100} + (-5) \times \frac{20}{100} + (-20) \times \frac{10}{100} + 100 \times \frac{40}{100} \\
&= 40.
\end{aligned}$$

Dados desbalanceados [1]

Há algumas alternativas para situações em que as classes estão desbalanceadas:

- Ajustar o modelo para maximizar a acurácia da classe minoritária;
- Escolher o corte para classificação com base na curva ROC;
- Poderar os dados com pesos maiores para as classes minoritárias;
- *Down-sampling*: amostra dados da classe majoritária para que tenha a mesma proporção da classe minoritária;
- *Up-sampling*: é feito um processo de reamostragem com reposição do grupo minoritário até tenha aproximadamente o mesmo número de observações que o grupo majoritário.

[1] Fonte: livro Applied predictive modeling

Resumindo...

- Seleção de modelos através de critérios como o C_p, AIC, BIC, R^2 e validação cruzada para o erro de previsão:
    - *best subset selection*;
    - *forward stepwise*;
    - *backward stepwise*.

- Métodos de encolhimento (*shrinkage methods*):

    - ridge;
    - LASSO;
    - elastic-net.
    - Modelo KNN para classificação.

- Regressão logística.

- Métricas para avaliar a performance de um modelo de classificação.

Obrigado!

magnotfs@insper.edu.br